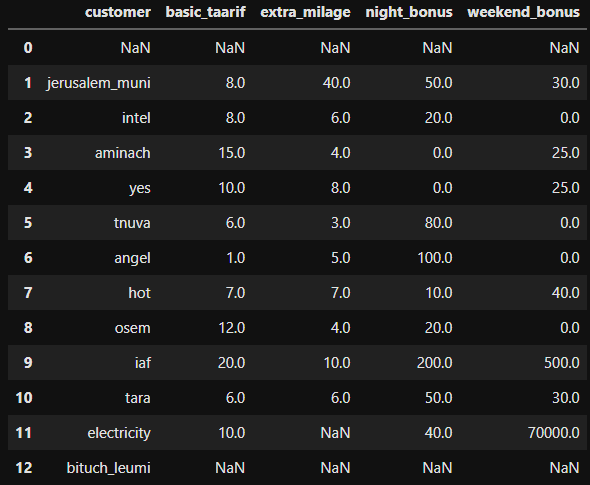
**Azrieli and Sons**

בתור התחלה עברתי על ניקוי הדאטה שקיבלתי מעזריאלי.

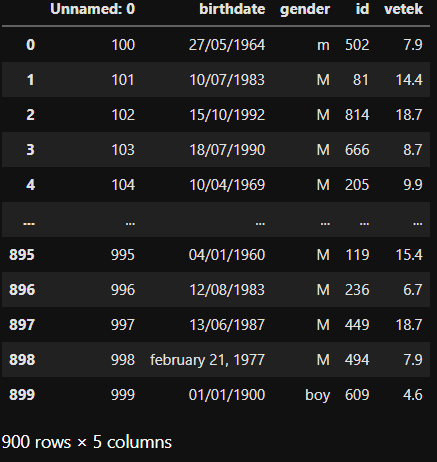
הדאטה של התעריפים של החברות:

אז שמתי לב שיש פה כמה NaN וביררתי עם עזריאלי לגבי חברת החשמל וביטוח לאומי והוא אמר

שהוא כבר לא עובד איתם. אז הצעתי למצע כל עמודה ולשים במקום הNan. ישנם ערכים גם שלא היו נראים הגיוניים (הבונוס בסוף שבוע של חברת החשמל) וגם במקומות האלה שמתי את הערך של הממוצע של העמודה המתאימה.

בנוסף, לאחר התוצאות הראשונות של השכר נראה שהמשכורות היו גבוהות ולכן ראיתי שהתעריפים של חיל האוויר וחברת מוני ירושלים הם גבוהים ביחס לממוצע וביררתי עם עזריאלי ומסתבר שהיה טעות של 0 בנתונים. לדוגמא, במקום 500 בבונוס סוף שבוע של חיל האוויר אמור להיות 50.

הדאטה של הנהגים החדשים והנהגים עם קביעות:



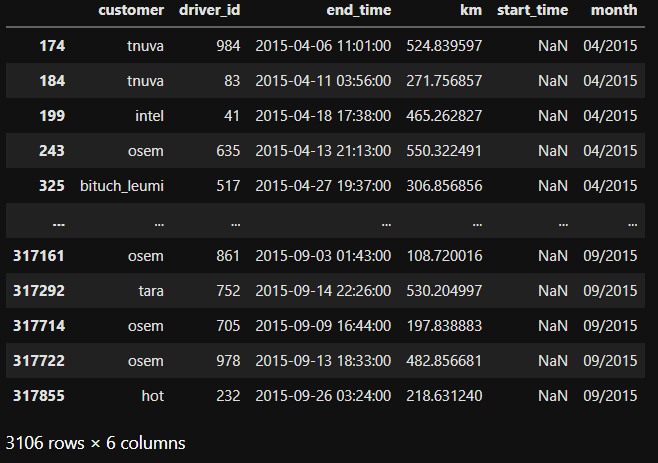
אפשר לראות מהתמונה וגם לאחר הסתכלות על הנתונים שהמגדר הוא לפעמים m, boy, none, male אז שמתי מגדר אחד לגברים M ולנשים F. בנוסף היה עוד כמה Nan בנתונים אז בגלל שיש רוב של גברים בדאטה אז אמרתי לעזריאלי שאני שם את הNan וnone ו Unknown בתור גברים.

בעמודת הימי הולדת ישנם גם כמה עמודות בלי ערכים, אז גם פה שמתי את הממוצע גילאים של כולם בערכים האלה שיצא שהם בני 50. כמה תאריכים היו גם חריגים כמו (1.1.1900) אז גם במקומם שמתי את הממוצע לאחר שעדכנתי את עזריאלי.

בנתונים של הנהגים החדשים בהתחלה היה מספר הימים בעמודה של הותק ואז שיניתי את זה לשנים.

הנתונים של הנסיעות:

תחילה שמתי לב שהסתכלתי על הזמנים של הנסיעות והיו לא מעט נסיעות שהיה להם Nan בתחילת הנסיעה או בסוף הנסיעה. אז עדכנתי את עזריאלי והפתרון שהצעתי הוא לחשב את המהירות הממוצעת של הנסיעות של העובד ולפי מרחק הנסיעה אז לחשב כמה שעות הוא נסע ובהתאם לזה לשים את הדלתא של הזמן של הנסיעה עם תחילת הנסיעה או סופה( תלוי איזה ערך היה חסר)

דוגמא לחוסר נתון של תחילת זמן בנסיעות:

בנוסף, לאחר חישוב השכר של העובדים בפעם הראשונה מצאתי שישנם נסיעות שאורכות יותר מ20 שעות. נתון שנשמע קצת לא הגיוני, ולכן החלפתי את הזמני נסיעות האלה בזמן הממוצע שהנהג נוהג בנסיעות. בהתאמה גם החלפתי את עמודת הקילומטרים בהתאמה.

מכיוון שיש את תעריף הextra\_milage( אחרי ה200 ק"מ התעריף הנ"ל מופעל). אז חישבתי עבור כל נסיעה שארכה יותר מ200 ק"מ את הזמן כאשר הנהג הגיע ל200 ק"מ. בהתאם לנתון הזה חישבתי את הזמנים של הסוף שבוע או לילה עד ל200 ק"מ( אם היו כאלה) וגם חישוב הזמן מה200 ק"מ עד סוף הנסיעה. לאחר מכן, הכפלתי את הזמן הזה במהירות הממוצעת של הנהג בנסיעה ובהתאם לכך חישבתי את הק"מ שהוא נהג לפני ואחרי ה200 ק"מ.

כדי לחשב את השכר הסופי של העובדים הייתי צריך לבדוק אם הנסיעות כללו זמן של סוף שבוע ואם הם כללו זמן של לילה. ( זמן של לילה = 22:00-06:00 , זמן של סופש= יום שישי 16:00-20:00 ביום שבת)

בנוסף, חישבתי עבור כל תת קבוצה של הדאטה( לדוגמא, נסיעות עם סופש וגם נסיעות של לילה וגם עברו את ה200 ק"מ) וכך הלאה.

כדי לחשב תעריף לילה או תעריף סופש אז הכפלתי את הק"מ שלפני ה200 \* התעריף הנ"ל באחוזים\* תעריף רגיל + ק"מ אחרי 200\* תעריף extra\_milage \* תעריף לילה/סופש. לזה מתווסף גם החישוב הרגיל של הנסיעה בלי קשר לתעריפים המיוחדים לדוגמא,

200\* תעריף רגיל + (km-200)\* תעריף אחרי 200. כלומר 200 ק"מ על התעריף הרגיל והק"מ שנותרו בתעריף של extra\_milage.

וכך חישבתי את השכר של כל נסיעה. הוספתי לכל עובד בהתאמה לחודש ומספר העובד שלו והוספתי בטבלת השכר.

לאחר החישוב של השכר הראשוני, אז בדקתי את הקורלציה של המחיר עם כל העמודות של הנתונים שיש לי. וכך הגעתי למסקנה שהנסיעות של חיל האוויר וירושלים מוני הן הכי יקרות לאחר הסתכלות של הנתונים של כמה כל חברה לוקחת תשלום עבור כל הנסיעות.

בנוסף, שמתי לב שישנם נסיעות ארוכות של 20 שעות ומעלה, שהתשלום עליהן גם היה גבוה ולכן הסרתי אותן והחלפתי בנסיעות שהן הזמן הממוצע של הנהג.