בתחילת התהליך ניסינו לחשוב מהי מחלקת ההיפותזות הכי הגיונית והנחנו שצריך להתחיל ברגרסיה לינארית. כמו בתרגיל 1 ניסינו לקחת את כל הפיצרים ולנסות לבדוק האם יש קורלציה בינהם לבין הלייבל שאותו ניסינו לחזות בעזרת פירסון קורליישן. בדיוק באותו אופן ניסינו להעלות בריבוע פיצרים, לכפול פיצרים אחד בשני (למשל מקדם ניפוח של אנשים שהמשיכו באנשים שהמשיכו) ולנסות לקבוע בין אילו פיצרים יש קורלציה חזקה מידי כדי להוריד אותם מהרגרסיה. גילינו שרוב הפיצרים נותנים קורלציות נמוכות מאוד, חלקן אפילו 0. הוקרלציה החזקה ביותר הייתה די טריוויאלית והיא בין מספר האנשים שהמשיכו לבין מספר האנשים שעלו. עוד קורלציה חזקה שהייתה היא בין מספר התחנה לבין מספר האנשים שעלו.

הדבר הבא שעשינו היה לנקות את הדאטא- באופן טריוויאלי הוא הכיל הרבה ערכי null וערכים לא הגיוניים, את clusters צריך היה להפוך מעברית למספרים ייחודיים לכל איזור.

ניסינו להוסיף עוד עמודות ששואבות מידע מהדאטא שלא נוכח באופן טריוויאלי אבל התגלה שרובן או נותנות קורלציה נמוכה יותר מהמקור (למשל להעלות בריבוע פיצר) או נותנות קורלציה 0.

בשלב הזה החלנו לעבוד עם 10% טריין. אחרי שסיימנו החלטנו שכדי להיות מסוגלים לבדוק שהקורלציות נשארות בערך אותו דבר העלו את כמות הטריין ל20% ולאחר מכן הקצנו 10% עבור טסט (בעצם dev).

בחלק השני הבנו שהמידע בתצורה הנוכחית לא מספיק טוב בתצורה הנוכחית. חישבנו את זמן הנסיעה, מספר התחנות, ואת המרחק הגיאוגרפי בין התחנות (בעזרת ספריה מתאימה של פייתון), מספר נוסעים ממוצע בין תחנות, ודחסנו את כל השורות שמכילות trip_id_unique להיות שורה אחת עם המידע המתאים לה.

חשבנו לחשב מרחק בין 2 תחנות ולראות קורלציה לכמות נוסעים שעלו\ זמן נסיעה אבל מפאת הזמן ומלאכת הסינון לא הספקנו להגיע לשם.

אחרי שמודל הרגרסיה של סעיף 1 עבד כמו שצריך ניסינו לחשוב על תהליכים נוספים שישפרו את השגיאה שאליה הגענו:

1. ללמוד בעזרת מודל אחד שעושה קלסיפיקסיה האם מדובר בשעת עומס או לא ולהכניס את המשתנה המציין הזה כעמודה (במקום לנסות לחשב לבד) ועל זה לבצע רגרסיה לינארית
2. יש דברים שלא מתנהגים בצורה לינארית- שעה ביום הוא גורם שמההבנה שלנו את העולם הוא דבר שמשפיע על כמות הנוסעים באוטוהוס אבל הוא לא לינארי ולכן חשבנו שאולי צריך לעבור למודל אחר. ניסינו להפעיל מודל (אנסמבל) שקראנו עליו באינטרנט שנקרא random forrest for linear regression. גילינו שעבור איטרציות של עצים בעומק 10,20 וריצות באורך 100,200 אנחנו מצליחים להוריד את השגיאה שלנו ב20% פחות מרגרסיה לינארית. הבעיה הייתה שהרצה כזאת לוקחת קצת יותר מ20 דקות והנחנו שזה לא יתאפשר בחוקי הפורמט.

בחלק 2 השתמשנו במודל בוסטינג של ספריה חיצונית כי לא היה מספיק קורלציות מעל 0.2 ולכן הרגשנו שמודל לינארי לא יעשה עבודה מספיק טובה.