Final Assignment - Predict Uber demand in New York

Economy in the Big Data World 2024

Dr. Roy Sasson; Ophir Betser

מגישים:

אריאל חדוות

איתן בקירוב

יובל בקירוב

30.10.24

הערות כלליות:

1. הקבצים המוגשים הם:
   a. מחברת ה markdown
   b. קבצי csv של נתונים חיצוניים
   c. דו"ח (מסמך זה)
   d. קבצי csv עם החיזויים
   e. קבצי csv נתונים
2. קישור לדרייב עם קבצי הדאטה הגולמי מהרשת וה- preprocessing ב- R שעשינו לו:
   https://drive.google.com/drive/folders/1eQ03fhgDG7U9rCFPy7cnrHuHZuSf0xbM?usp=sharing
3. עבור הויזואליזציות שצירפנו: בדר"כ, צבע כחול בהיר משתייך לדאטה המכיל מרחקים של עד 1000 מטר רדיוס מהאמפייר סטייט וצבע ירוק בהיר משתייך לדאטה המכיל מרחקים של מעל 2000 מטר רדיוס מהאמפייר סטייט.

   כאשר בצד ימין הויזואליזציה שייכת לדאטה ה-1000 ומצד שמאל עבור דאטה ה- 2000.

Section A: Data Rearrangement

נתוני האימון שלנו שונים מאלו של המבחן. הסרנו פיצ'רים לא רלוונטים, סידרנו לזמני אינטרוולים של 15 דקות, הסרנו דאטה לא רלוונטי ועוד:

עבור train_raw_data, הסרנו רשומות שמציגות נסיעות מעל 1000 מטר מהאמפייר סטייט ע"י **סינון מרחק**. נקרא לדאטה המסונן - train_filtered_1000m_data.

בנוסף, ביצענו **סינון זמנים**. הסרנו רשומות מ-2 הדטאות (train_filtered_1000m_data ו-train_filtered_2000m_data שמתייחס לרשומות המייצגות נסיעות מעל רדיוס של 2000 מטר מהאמפייר סטייט) המכילות נסיעות שהתבצעו בזמנים שהם לא בין 18 ל-23:45 (בדומה לטסט). בנוסף, יצרנו **אינטרוולי זמן** של 15 דקות, ווידאנו תקינות התאריכים, תקינות שעות ותקינות קודי ה base שמייצגים קוד חברה עבור האובר. בנוסף, ראינו שאין NAs.

לאחר מכן, **הסרנו את הפיצ'רים המיותרים** ונשארנו (עבור 2 הדטאות) עם: time_interval ו-number_of_pickups. שם ווידאנו שאין **אינטרוולי זמן חסרים** (כלומר שהיו בהם 0 נסיעות אובר), הוספנו במידה והיה חסר עם number_of_pickups = 0, על מנת לא להרוס את רצף ה- time series.

Section B: Telling story with the data

תחילה, **הוספנו פיצ'רים מתוך הנתונים החדשים** שיש לנו: **is_weekend** - שמבחין בין נסיעה ביום חול (=0) ויום שהוא סופ"ש (=1, ראשון ושבת בארה"ב). זוהי תכונה חשובה מכיוון שדפוסי האיסוף לרוב משתנים בימי חול לעומת סופי שבוע. בנוסף, **is_night** - שמזהה אם מרווח הזמן הוא בערב או בלילה. במקרה שלנו, ערב (=0 אם: 18:00 - 20:30 כולל) ולילה (=1 אם: 20:31 - 23:45 כולל). בנוסף, הוספנו את קטגוריות הזמן בהתאם ל-time_interval (**פיצלנו לדקות, שעות, חודשים וימים**),נציין כי המודל לא יכול להתאמן על הפורמט שה-time_interval נמצא בו כרגע.

לאחר מכן, **הוספנו תכונות מהרשת** (preprocessing עבורם נמצא בקבצי R נפרדים בקישור המצורף ב"הערות כלליות"):

- **מזג אוויר (Weather)**- מכיל תכונות טמפרטורה, לחות, מהירות הרוח, "feels like", עננות, האם ירד גשם בשעה האחרונה (בינארי) והאם ירד שלג בשעה האחרונה (בינארי) באזור האמפייר סטייט. חקרנו כי מזג האוויר באזור בניין האמפייר סטייט דומה לממוצע בניו יורק, עם הבדלים קלים כמו טמפרטורות גבוהות יותר ורוחות חזקות יותר בשל הסביבה העירונית, לכן הוספנו נתונים אלו לשתי הטבלאות (כרגע).
- **נתוני תאונות (Crashes)** - מכיל מספר אנשים שנפצעו, מספר אנשים שנהרגו, מספר תאונות. מידע שעשוי לעזור לנו להבין כיצד תאונות דרכים עשויות להשפיע על הביקוש של אובר.
- **נתונים על אירועים מורשים בניו-יורק (Events)** - מכיל תכונת "מספר אירועים" שמונה עבור כל מרווח זמן, כמה אירועים קיימים בכל NYC. מידע שעשוי להשפיע על התחבורה במידה ויש ימים עמוסי אירועים.
- **נתוני מוניות צהובות (Yellow Taxis)**- מכיל את כמות ה pickups עבור המוניות הצהובות עבור אינטרוולי זמן של 15 דקות עבור המרחק המתאים פר דאטה. חשבנו שמדובר בנתון מאוד מעניין שעשוי להראות קורלציה (חיובית או שלילית) בין נסיעות במוניות לנסיעות באובר.
- **נתוני מעצרים (Arrests)** - מכיל מידע על מעצרים שבוצעו באזור המתאים ומכיל בתוכו את סה"כ מעצרים (כולל מעצרים הקשורים לתחבורה וכאלה שלא).
- **נתונים על חגים (Holidays) -** שילבנו טבלה של כל החגים הפדרליים בארה"ב עם התאריכים שלהם בשנת 2014 . חגים מובילים בדרך כלל לשינויים בדפוסי הנסיעה לעבודה, מכיוון שפחות אנשים נוסעים לעבודה

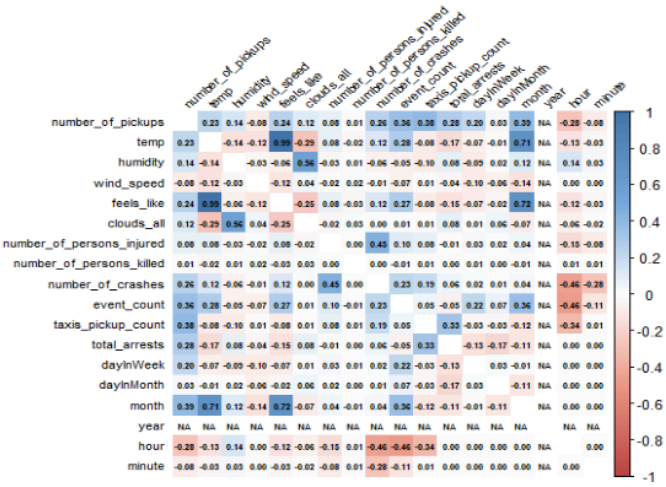
ויש פחות נסיעות ויותר זמן משפחתי, מה שעלול להקטין את הביקוש לנסיעות באובר. (החגים בתאריכים שלנו הם: Memorial Day, Independence Day, Labor Day).

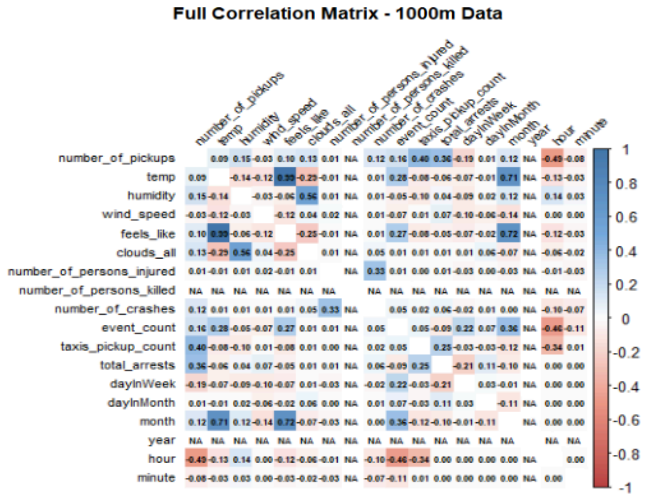
לאחר הוספת כל התכונות - הדאטה נקי, ללא ערכים חסרים וללא כפילויות.

בהמשך נעבור על כל התכונות וננקה את אלו שנראות לנו רלוונטיות פחות (כדי שלא יהיה מצב של ממדיות יתר שיכול להוות בעיה בביצועים של המודלים שלנו, ככל שהממדיות גדולה יותר יש יותר "רעש" ואנחנו יכולים להגיע לאוברפיטינג. בנוסף ככל שיש יותר ממדים, קשה יותר לזהות כל מיני דפוסים בדאטה שלנו, מה שיכול להקשות על אימון המודל או על דיוקו).

בחלק של ה- **Exploratory Analysis** הסתכלנו על הפיצ'רים של הדאטה שלנו וביצענו ויזואליזציות על מנת ללמוד את הדאטה שלנו טוב יותר ואת הקשר למשתנה המוסבר (number_of_pickups).

מהסתכלות על **correlation matrix** של המשתנים הנומריים:

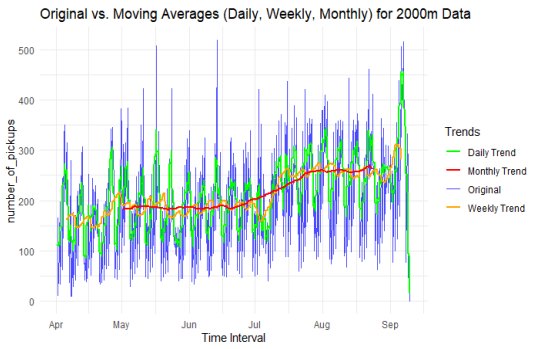


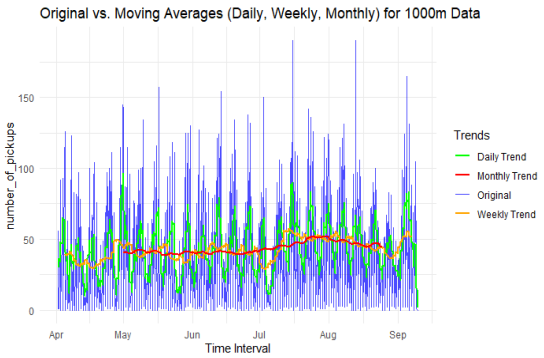


1. feels_like ו- temp עם קורולציה של 0.99, כמעט מושלמת. לכן נסיר אחד מהם כדי למנוע מצב של מולטיקוליניאריות (ווידאנו זאת גם עם מדד VIF). הסרנו את temp כי ל- feels_like קורולציה גבוה יותר עם המשתנה המוסבר.
2. מופעי ה-NA שמופיעים ב- number_of_persons_killed ו year מעידים כי משתנים אלו מכילים לכל אורך הדאטה רק ערך אחד (לאחר בדיקה, number_of_persons_killed מכיל רק 0ים ו year מכיל רק 2014). לאחר בדיקה במערך הנתונים של ה - 2000 מטר ומעלה, נראה ש- number_of_persons_killed מכיל מעט מאוד ערכים שהם לא 0 והוא פחות רלוונטי, לכן נסיר אותו גם מפה.
3. נשמיט את המשתנים wind_speed, humidity ו - clouds_all בגלל קורלציה נמוכה עם המשתנה המוסבר, ומכיוון שמשתנים אלו לא עשויים להשפיע בצורה ישירה מובהקת על המשתנה המוסבר בהשוואה למשתני הטמפרטורה האחרים.

בגרפי ה- boxplot שביצענו ראינו שהמשתנה is_snowing_last_hour קבוע על 0, כיוון שלא היה שלג בין אפריל לאמצע ספטמבר. מאחר שהוא לא תורם, הסרנו אותו.

טרנדים - בעזרת ממוצע נע נסתכל על המגמה על פני זמנים שונים עבור הביקוש לנסיעות Uber ב-2 הדטאות (מימין 1000, משמאל 2000):

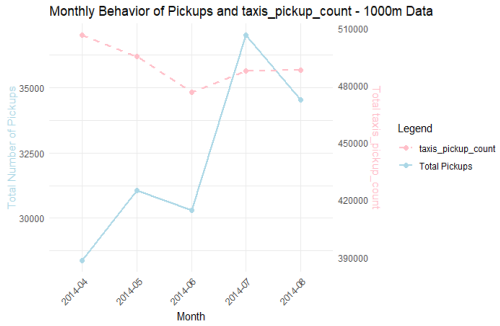


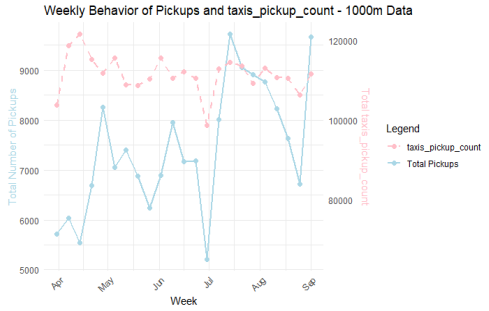


הביקוש ל-Uber באזור האמפייר סטייט הוא דינמי, עם מגמת עלייה בביקוש לאורך הזמן, במיוחד ברדיוס של 2000+ מטר הכולל אזורי תיירות, עסקים ומגורים. במגמות היומיות והשבועיות נראית תנודתיות המושפעת כנראה מגורמים כמו מזג אוויר, תאונות, חגים וסופי שבוע. התקופות עם הביקוש הגבוה ביותר הן חודשי הקיץ (יולי ואוגוסט), עם ירידות קטנות עקב אירועים מיוחדים חוזרים. לאחר סקירה, מצאנו כי מרבית אירועים אלו משקפים את הדינמיקה העירונית של ניו יורק ולכן השארנו אותם בנתונים. בנוסף, קיימים משתנים לייצוג חגים, סופי שבוע וסה"כ אירועים מיוחדים.

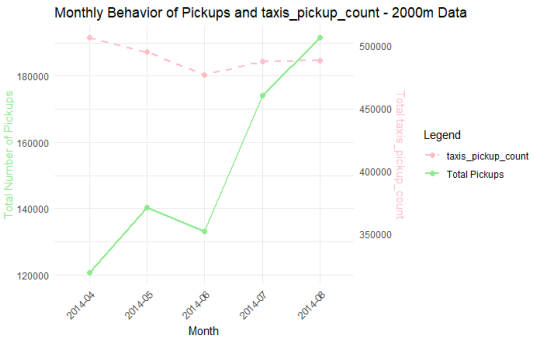
התנהגות Uber לעומת מוניות: למרות שהמוניות מציגות שינויים מתונים יותר (בשל הפופולריות והביסוס שלהן בניו יורק ב-2014), שני השירותים נוטים לעקוב אחר מגמה דומה בביקוש. זה מצביע על כך ש-Uber ומוניות מגיבים דומה למחזורי ביקוש רחבים בעיר, למרות הבדלים יומיומיים. ניתן לראות זאת בגרפים:

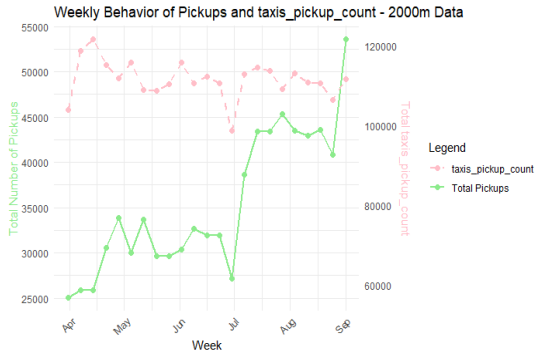
עבור ה- 1000:





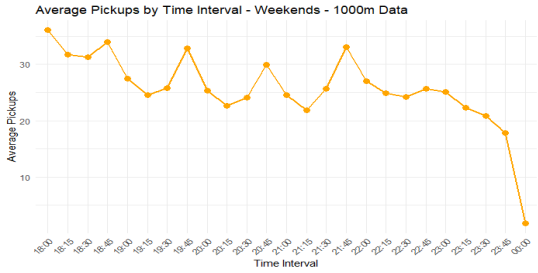
עבור ה- 2000:

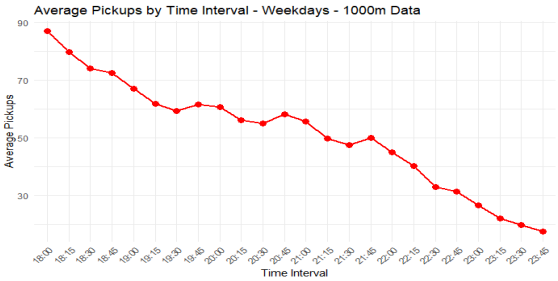




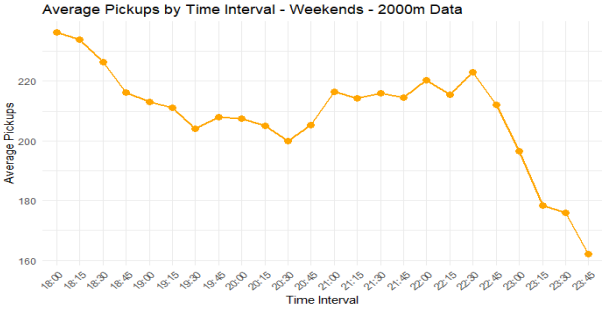
בהסתכלות על **ימי חול לעומת סופש (שבת וראשון)** ראינו כי עבור 2 מערכי הנתונים יש ירידה מגמתית בביקוש לאובר בימי חול (גרפים אדומים) ככל שהזמן עובר (כנראה כי אנשים חוזרים הביתה מהעבודה בשעות הערב המוקדמות). הגרף האדום עבור דאטה +2000 מראה ירידה מתונה יותר וקצת יותר יציב, ככל הנראה כי מכיל שטח גדול יותר ל- pickups. עבור סופי שבוע (גרפים כתומים), דאטה ה- 1000 מראה קצת יותר תנודתיות כי ייתכן כי קיימים פחות אזורי תיירות בתחום זה ביחס לשטח של בדאטה ה2000+. בדאטה ה-2000 הביקוש מראה פחות פיקים, להשערתינו כי מכיל יותר מאזור תיירות מרכזי אחד כך שה- pickups מתבצעים לתקופות זמן ממושכות יותר.

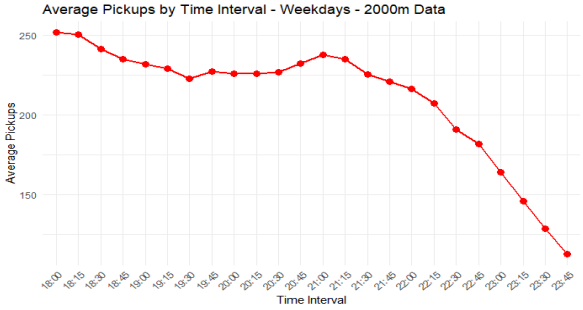
1000:



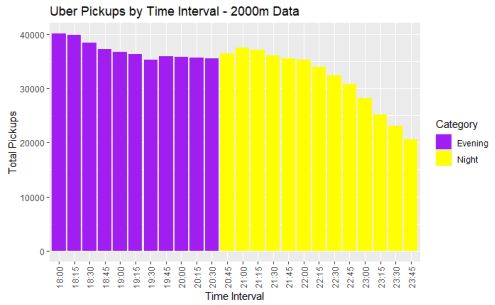


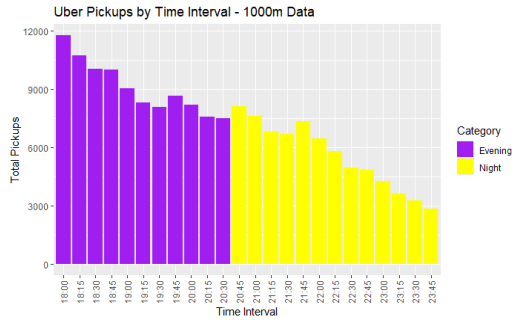
2000:





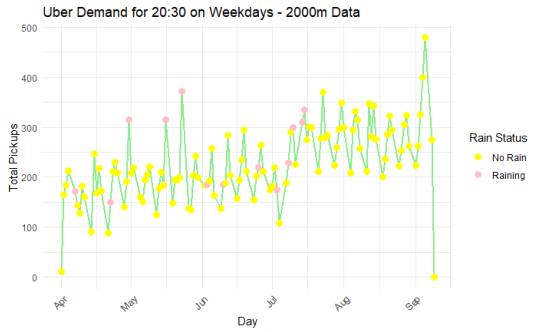
בהסתכלות על **שעות ועל זמני ערב ולילה,** כאשר הסתכלנו על סה"כ pickups פר רבע שעה (אינטרוול), ראינו שיש יותר pickups בערב מבלילה - רואים זאת בבירור בדאטה ה- 1000, בדאטה ה- 2000+ רואים ירידה פחות חדה בלילה בהשוואה לירידה בדאטה ה-1000 - כנראה בשל בילויי הלילה שיש בניו יורק שבאים בטווחים אלו יותר לידי ביטוי. (מצד ימין 1000, מצד שמאל 2000. כאשר סגול זה ערב וצהוב זה לילה).

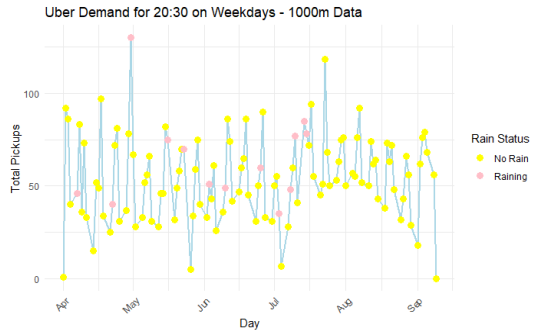




בנוגע ל**מעצרים** עושה רושם כי קיים מתאם בין מספר המעצרים לביקוש ל-Uber. ייתכן שחלק מהמעצרים קשורים לתקריות תנועה כמו נהיגה פזיזה או תאונות "פגע וברח" המשפיעות על זרימת התנועה. כאשר כבישים חסומים או יש עיקופים, אנשים עשויים להעדיף Uber. מצד שני, אם התקריות גורמות לסגירת רחובות מרכזיים, הדבר עלול דווקא להפחית את הביקוש ולהוביל לשימוש באפשרויות אחרות כמו הליכה או רכבת תחתית (הגרפים הרלוונטים במחברת, עקב חוסר במקום).

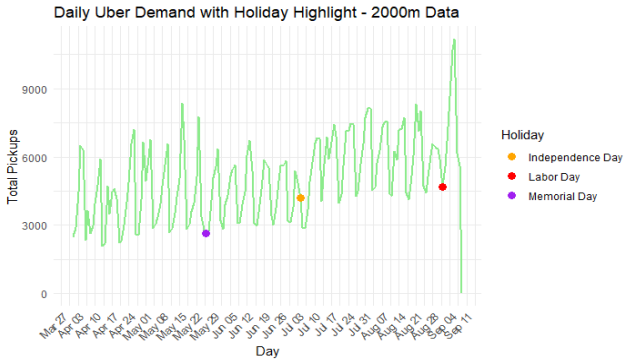
בנוגע להשפעת ה**גשם** הסתכלנו על שעות מסוימות בדאטה ועל הביקוש לאובר במשך תקופת הזמן, תוך סימון האם ירד או לא ירד גשם. נצרף דוגמה עבור השעה 20:30 (מימין ה-1000 ומשמאל ה- 2000 כאשר ורוד הם יום בו השעה גשומה וצהוב כשלא גשום):

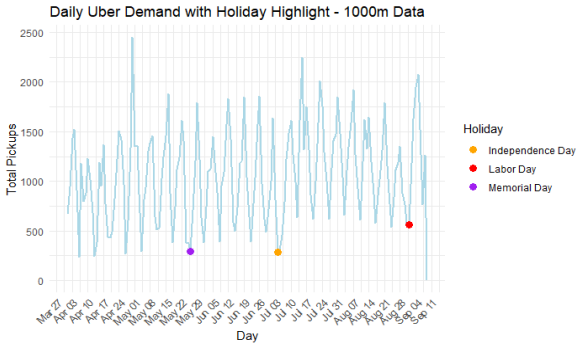




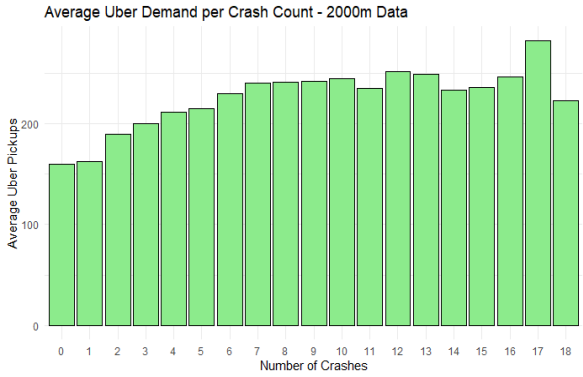
בשעות השיא (18:00 ברדיוס 1000 מ' וה- 2000 מ'), הביקוש ל-Uber נותר גבוה לרוב ולא מושפע מגשם, כנראה בשל הצורך בנסיעה הביתה או לעבודה. הסתכלנו גם על שעות הערב-לילה (אבל לא מאוחר מידיי כמו: 20:30 ו- 22:30), כדי להבין טוב יותר את השפעת הגשם. נראה שבימי גשם הביקוש לאובר עולה - אנשים מעדיפים להימנע מהרטבות. בערבים מאוחרים, השפעת הגשם משתנה, השערתינו היא שחלק יישארו בבית בגלל הגשם (ולכן רואים ירידה בביקוש לאובר), אך מי שיוצא יעדיף Uber על פני האלטרנטיבות.

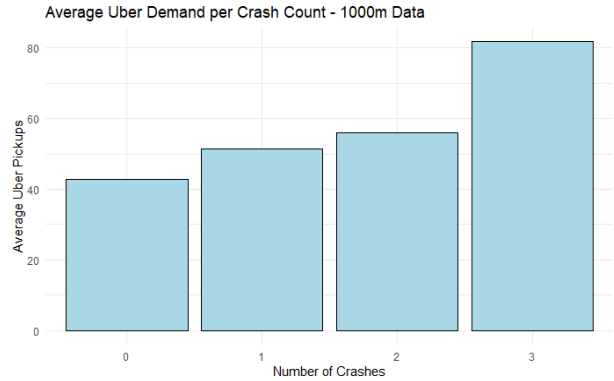
כשהסתכלנו על **חגים,** זיהינו דפוס עקבי של ביקוש נמוך לUber בשלושת החגים (**Labor Day, Independence Day ו**-**Memorial Day**) בהשוואה לימים שאינם חג - איחדנו אותם לתכונה אחת, is_holiday. מגמה זו של ביקוש נמוך בחגים הגיונית, כי טבעי שבתקופת החגים פחות אנשים נוסעים לעבודה ומעדיפים להישאר עם משפחה. (מימין 1000 ומשמאל 2000):





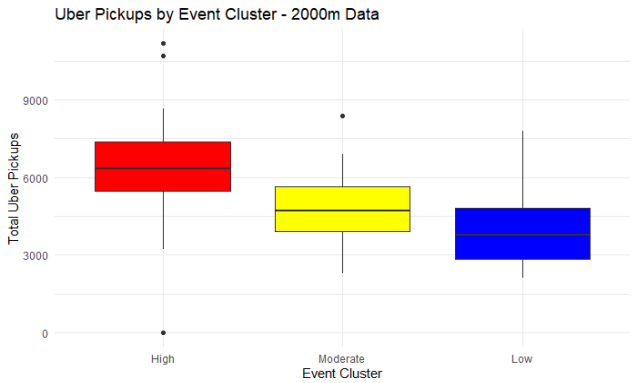
בהסתכלות על **תאונות ופציעות**, עבור 2 מערכי הנתונים, כמות האנשים שנפצעו פר 15 דקות היא לרוב 0, לכן ראינו לנכון להפוך את הפיצ'ר number_of_persons_injured לפיצ'ר בינארי המייצג האם בן אדם נפצע או לא (is_persons_injured). בנוסף, הסתכלנו על number_of_crashes ועל הקשר למשתנה המוסבר:

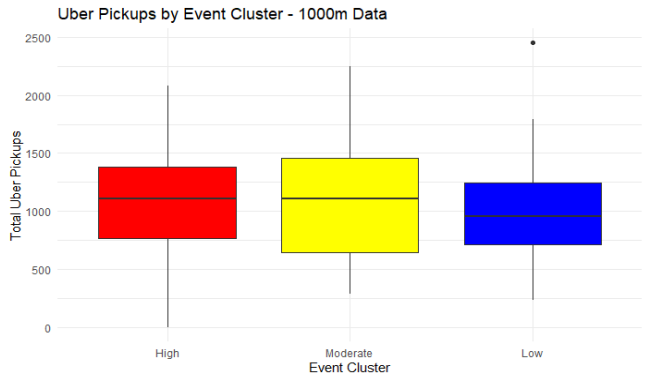




רואים שישנו קשר חיובי בין מס' תאונות למס' הpickups (לרוב). הגיוני לומר כי ככל שיש יותר נסיעות אובר, יש יותר תנועה בכבישים ולכן גם יותר תאונות. בנוסף, סביר שמספר התאונות בדאטה 1000 נמוך יותר בהשוואה לזה של ה 2000+, שכן ישנם פחות כבישים באיזור.

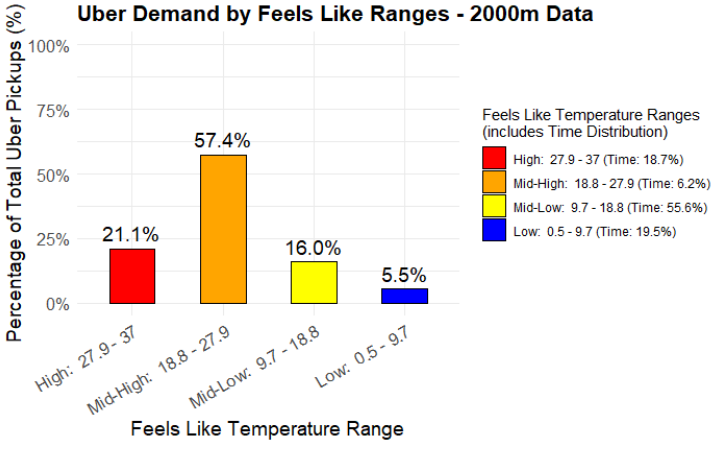
עבור **סה"כ האירועים בניו-יורק**, ברדיוס של 1000 מטר מהאמפייר סטייט, מספר האירועים בעיר משפיע מעט מאוד על ביקוש האובר, ללא הבדלים משמעותיים בין מקבצי אירועים "נמוכים", "בינוניים" ו"גבוהים". לעומת זאת, ברדיוס של 2000 מטר, אירועים רבים יותר מצביעים על עלייה בביקוש לאובר, מה שמצדיק שמירה על תכונת total_events במערך הנתונים של 2000 מטר והסרתה ממערך ה-1000 מטר, שבו השפעתה כמעט ואינה מורגשת.

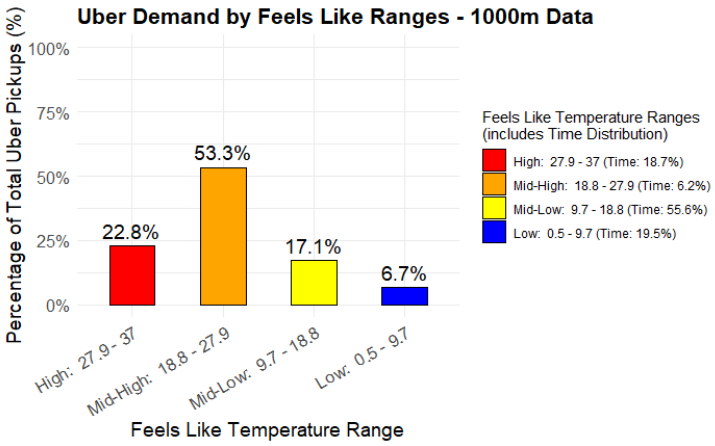




בנתון ה **feels like** חילקנו את הדאטה לארבע קבוצות: feels like גבוה עד נמוך. ראינו שהביקוש לאובר הוא הגבוה ביותר בטמפ' feels like נוחה (18.8°C - 27.9°C), כאשר אנשים נוטים לצאת יותר במזג אוויר נוח ולכן גם לנסוע יותר באובר. הביקוש יורד בתנאים קיצוניים - חמים (27.9°C - 37°C) וקרים (0.5°C - 9.7°C) - יכול להעיד לדעתינו על כך שאנשים מעדיפים להישאר בבית כשמזג האוויר לא נוח (ולעבוד מהבית למשל).

(מימין 1000, משמאל 2000):





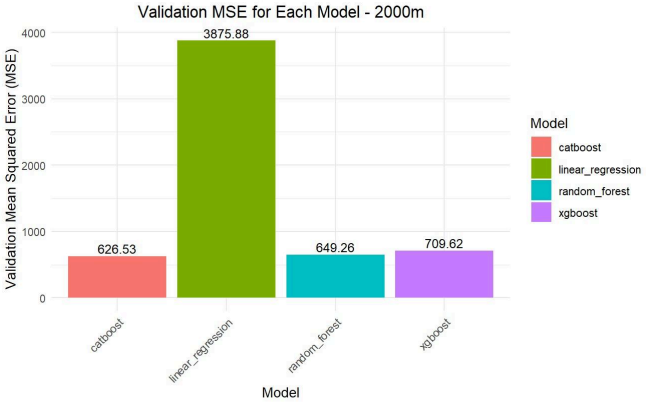
Preprocessing קצר לפני שנקפוץ למודלים: וידאנו שכל הפיצארים עם ה data typeים הנכונים, הסרנו את פיצ'ר ה- time_interval והסרנו את ה- pickups ששווים לאפס מאחר ולאחר בדיקה הוא מוריד את תוצאות המודל.

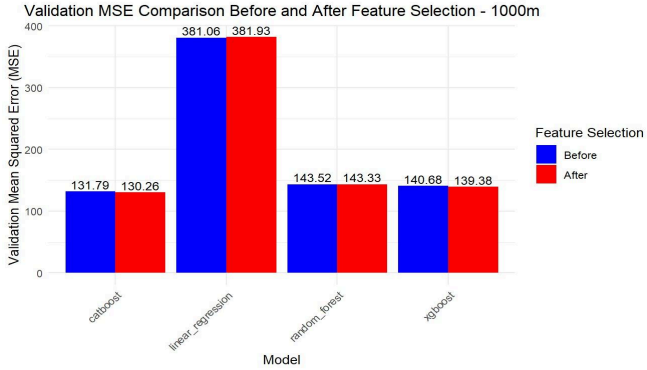
Section C: Forecast for the future

בדקנו מספר מודלים: linear_regression, random_forest, xgboost, catboost כאשר לכל אחד מהם יתרונות וחסרונות ולבסוף בחרנו את המודל עם ציון MSE הנמוך ביותר לצורך החיזוי עבור נתוני המבחן.

linear_regression- מודל פשוט יחסית שמניח קשר ליניארי בין הפיצארים למשתנה המוסבר; **Random Forest**- מודל שמשתמש בהרבה עצים כדי להגיע לתחזיות מדויקות יותר, ועושה ממוצע של התוצאות שלהם. הוא טוב בהתמודדות עם רעש ומפחית את הסיכון לאובר-פיטינג; **XGBoost**- מודל boosting מהיר ואפקטיבי, שמשפר את הביצועים על ידי תיקון טעויות הדרגתית בכל עץ חדש; **CatBoost**- מודל boosting שמצטיין בטיפול בתכונות קטגוריות ובביצועים מהירים, מתאים במיוחד לנתונים עם תכונות קטגוריות רבות.

בכל המודלים השתמשנו בשיטות **K-fold cross-validation** ו-**grid search** כדי להעריך את ביצועי המודל באופן אמין ולהקטין את הסיכוי לאובר-פיטינג. **K-fold cross-validation** מאמן ובודק את המודל על חלקים שונים של הנתונים, מה שמונע התאמה יתרה לרעש של חלק מסוים בלבד. בנוסף שיטה זו מתאימה לדאטה סט יחסית קטן. **Grid search** מאפשר למצוא את הפרמטרים הטובים ביותר למודל, תוך חיפוש מקיף ובחינת שילובים שונים של פרמטרים, מה שעוזר לשפר את ביצועי המודל ומניעת overfitting. בנוסף, עבור דאטה ה-1000 השתמשנו ב- **Feature Importance** שבאמצעותו יכולנו לראות את דירוג הפיצ'רים לפי מידת התרומה שלהם לביצועי המודל, כך שאפשר להבין אילו תכונות חשובות יותר לתוצאה הסופית ואילו פחות. ע"י הסרת תכונות פחות רלוונטיות (is_night , number_of_crashes , is_persons_injured) עזרנו למודל להיות פשוט יותר וממוקד, מה שמפחית את הסיכוי ל- overfitting עם סט האימון וגם כך הורדנו (כמעט לכל המודלים) את ציון השגיאה. עבור דאטה ה-2000 פעלנו אחרת, החלטנו לבסוף להסיר את הפיצ'רים שקשורים למזג האוויר (, feels_like is_raining_last_hour) היות והם מייצגים את מה שקורה באמפייר סטייט ולא בהכרח מצייגים את מה שקורה מעבר לכך, דבר שעשוי לשבש את ביצועי המודל. החלטנו להשאיר בדאטה ה- 2000+ רק פיצ'רים שיודעים להבדיל בין ההתנהגות של ה-1000 וה-2000.





כפי שניתן לראות, עבור הדאטה שמתייחס לביקוש נסיעות אובר בתוך רדיוס ה-1000 מ' מהאמפייר סטייט, המודל שעבד הכי טוב עבורו היה: **catboost** עם ציון MSE של: 130.26.

עבור הדאטה שמתייחס לביקוש נסיעות אובר מעבר לרדיוס ה-2000 מ' מהאמפייר סטייט, המודל שעבד הכי טוב עבורו היה (גם): **catboost** עם ציון MSE של: 626.53.

לאחר חשיבה והסתכלות נוספת על דאטה ה-2000, ניתן היה לשפר את התחזיות אם היינו בוחרים מיקום כלשהו בדאטה שדומה לאמפייר סטייט ולהסתכל על הביקוש לאובר בתוך רדיוס ה-1000 מ' ממיקום זה ולהשתמש בדאטה הזה בלבד לצורך אימון המודל. יכול להיות שזה היה מדייק את ביצועי המודל בנוסף לבחירה חכמה של פיצ'רים (כמו שעשינו כחלק מעבודת ה- data engineering המעמיקה שעשינו לאחר כל העבודה).