Flow

Data Preprocessing

החלק הראשון והכי מעצבן ותכלס קשה בפרויקט שלנו היה להשיג את הנתונים וליצור משהו שיהיה ניתן בכלל לעבוד איתו. את הנתונים של המזהמים לקחנו מדאטה בייס אחד ואת הנתונים של המזג אוויר מאתר אחר. בצורה ידנית חילצנו קובץ CSV לכל אחד מ4 המזהמים עבור השנים 2016, 2018, 2020 ו 2021. כמו כן קובץ אחד של מזג אוויר בעיר ניו יורק סיטי בשנים 2016-2022.

הבעיה בקובץ של המזהמים הוא שיש שמה לכל שנה הרבה יותר רשומות (יותר מ365) – הסיבה לכך זה שיש הרבה מחוזות במדינת ניו-יורק. ההנחה שלנו לפרויקט זה שהמזג אוויר בכל המחוזות שקרובים לניו יורק סיטי זהה (מחוזות כמו Bronx, Queens יחשבו קרובות לניו יורק סיטי כדי להניח שהמזג אוויר בהם זהה. לעומת זאת מחוזות כמו Erie רחוקים בצורה משמעותית כך שלא נניח זאת לגביהם וגם בהמשך נוריד אותם מהקובץ נתונים שלנו). אחרי שהשגנו את הנתונים עבור המזהמים נבצע preprocessing כדי לחלץ קובץ אחד. הפיצ'רים שהשארנו מהקבצים של המזהמים (עבור כל מזהם) הינם : [Date, pollutants (e.g., Ozone, NO ..), AQI (i.e., Air Quality Index), County, Site Latitude, Site Longitude]. אחרי שביצענו merge לכל הדאטה סטים של כל מזהם בנפרד כדי לקבל קובץ לכל מזהם עבור השנים שאני בוחנים, ביצענו join בין כל טבלאות הנתונים על הפיצ'רים ['Date', 'County', 'Site Latitude', 'Site Longitude'] -> כך שהטבלה הסופית של המזהמים כוללת 2896 שורות ו12 עמודות (פיצ'רים) בחלק זה. הדבר החשוב ששווה לשים לב זה שנשארו לנו שתי מחוזות בקובץ נתונים לאחר שלב זה -> Bronx, Queens, אלו מחוזות שנמצאים ממש בNYC, כך שהקובץ שלנו לא מכיל מידע על מחוזות מרוחקים שלא רלוונטים עבורינו. אחרי שסיימנו לטפל בקובץ של המזהמים עברנו לקובץ של המזג אוויר. בקובץ זה כמעט ולא היה מה לגעת בחלק הראשון. החלק האחרון בשלב הpreprocessing כלל join בין הטבלה של המזהמים וטבלה של המזג אוויר על העמודה של Date. בסיום השלב הזה נשארנו עם קובץ שמכיל 2896 שורות ו21 עמודות (פיצ'רים) לשלב זה. בשלב זה טיפלנו גם בערכים חסרים (רק בקובץ של המזג אוויר היו סהכ 173 שורות עם ערכים חסרים, הורדנו אותם מהדאטה שלנו)

EDA (exploratory data analysis)

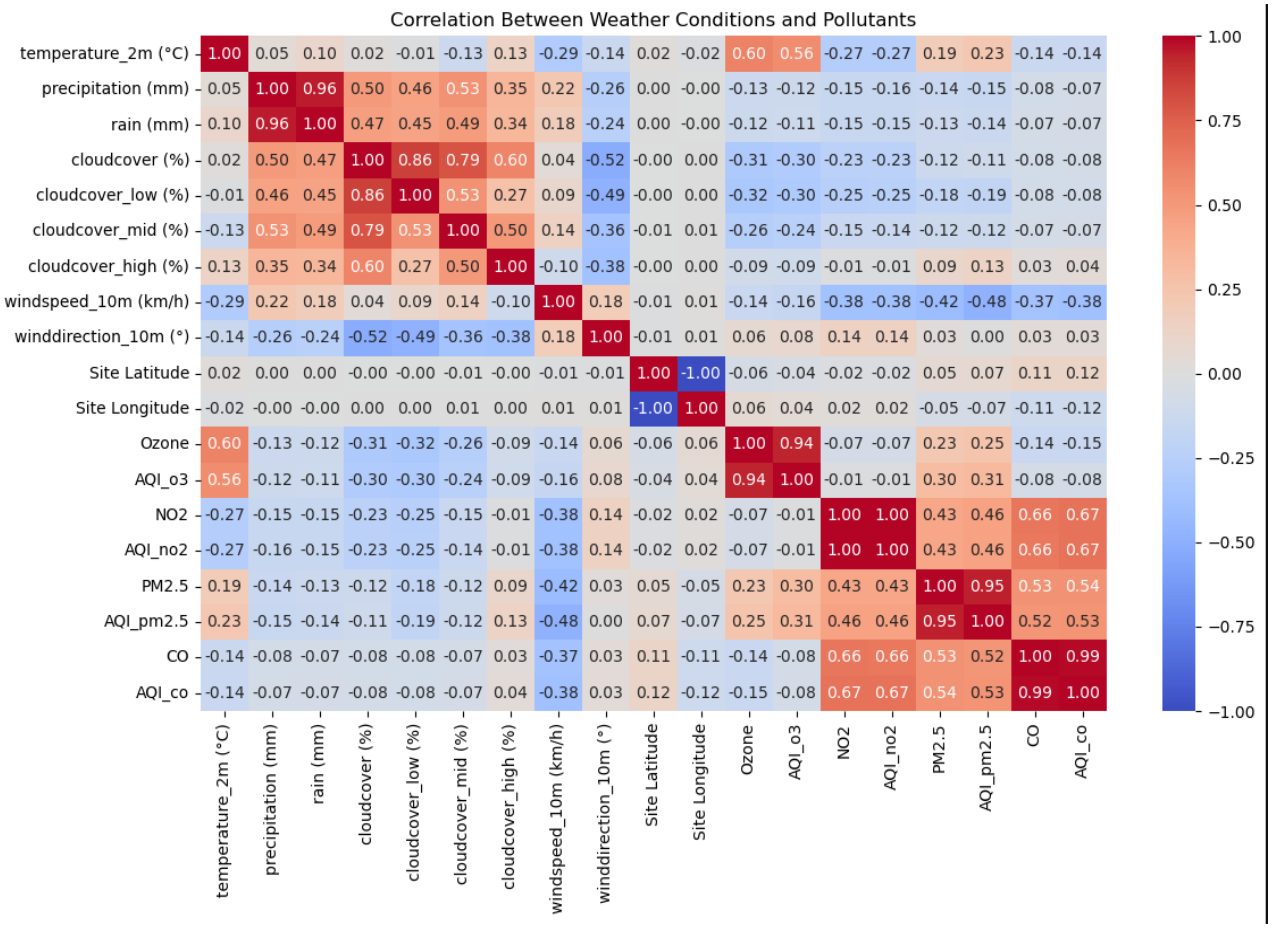
Data Visualization & Conclusions

בחלק זה ביצענו מספר שיטות ויזואליזציה ע"מ להבין איך הנתונים שלנו מתפלגים, קורלציה בין משתנים, וכן השפעה בין מזהמים על המזג אוויר.

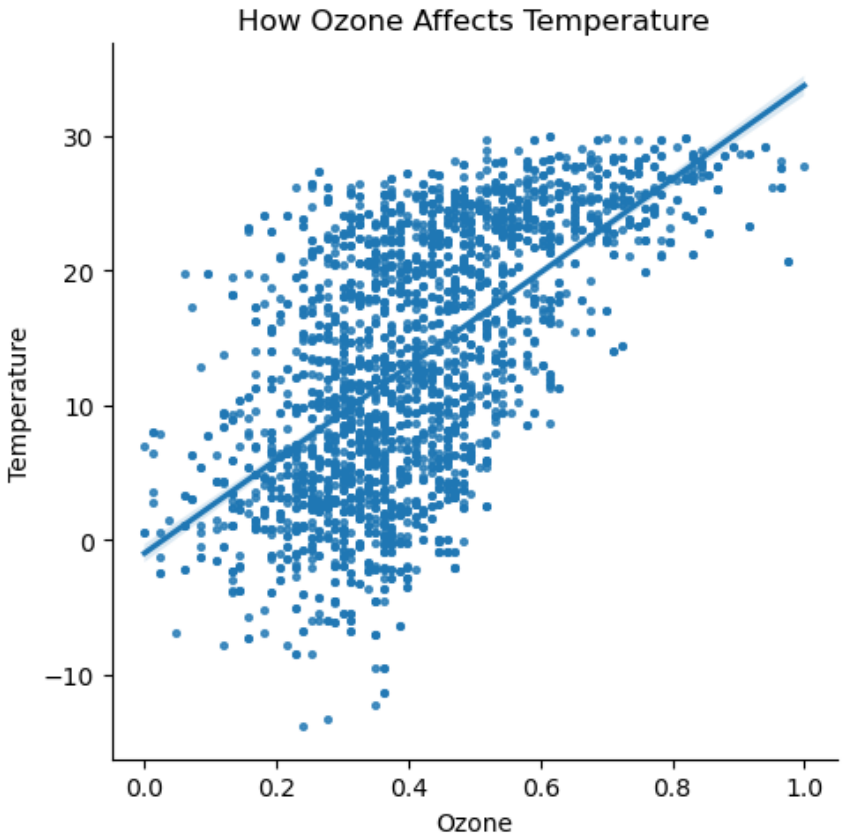
- ע"י גרף היסטוגרמה ראינו כי רוב הפיצ'רים שלנו מתפלגים נורמלית (בין אם נורמלי רגיל או זנב ימין/שמאל). עם זאת, בחרנו לנרמל את הנתונים שלנו ע"י MIX MAX SCALER ולא ע"י Normal Scaler וזה כי רצינו שכל הנתנוים שלנו יהיו בטווח של 0-1 מכמה סיבות:
  - ויזואליזיציה – יצרנו אחר"כ עוד גרפים שיותר נוח להבין ולהסיק דברים כאשר כל הערכים שלנו חיוביים.
  - Feature engineering – בשלב זה שקורה מיד לאחר הEDA, יצרנו פיצרים ע"י ONE HOT ENCODING גם לDATE, גם לCounty. כאשר למשל רשומה עם תאריך – 14/1/2018 קיבלה את הערך הבא עבור YEAR -> [0 0 1 0] (כאשר הסדר זה 2022, 2020, 2018, 2016) – כך שכל הדאטה שלנו ישאר מנורמל בין 0-1.
  - המודל שאנו בונים הינו מודל שמבוסס על רשת ניורונים – מודל MLP פשוט – נעדיף שהמודל יקבל וקטורי EMBEDDINGS שמכיל ערכים חיוביים, כך יוכל ללמוד לתת יחס חשוב יותר לערכים שקרובים ל1 ויחס פחות חשוב לערכים שקרובים ל0.
- גרף קורלציה בין משתנים: דברים שמאוד תפסו לנו את העין זה שיש קורלציה חיובית בין אוזון (OZONE) לבין טמפ' (0.56, 0.6) – כך ששווה לבדוק אותם בנפרד ולראות איך עליה ברמת האוזון משפיעה על עליה בטמפ. עוד דבר מעיין ששמנו לב זה שיש קורלציה חזקה בין הLABELS גשם

ומשקעים יחד עם רוב הפיצ'רים שקשורים למזג אוויר – דבר זה פחות טוב למחקר שלנו משתי סיבות:

- יש multicollinearity בין שתי הLABELS גשם ומשקעים (כמעט 1 בשתיהם) – דבר שיכול מעוד להשפיע על החיזוי של המודל שלנו אם נבחר לחזות את אחד מהם.
- ההשערה ומה שאנחנו הכי מעוניינים במחקר שלנו זה לבדוק איך המזהמים והשינוי שלהם משפיע על אלמנטים של מזג אוויר – וכיוון שיש הרבה פיצ'רים שנמצאים בקורלציה חזרה (כל הכתום בצד שמאל למעלה) עם הLABELS גשם ומשקעים, אם נבחר לאמן מודל ולחזות אותם, יתכן שרוב התוצאות שנקבל יהיו בגלל הקורלציה החזקה בין אותם פיצ'רים של מזג האוויר (לגשם+משקעים) -> זה גם אגב אחת הסיבות שבחרנו בתהליך האימון לא להכליל את גשם ומשקעים בתור LABELS וניסינו לחזות רק את temperature שנמצאת בקורלציה חלשה עם רוב הפיצ'רים של מזג האוויר.
- נשים לב שיש קורלציה חזקה בין המזהמים לבין עצמם – שזה הגיוני (מעיין בדיקת שפיות) – אני ציפינו שכאשר מזהם אחד עולה גם הרמה של מזהם שני יעלה, וההפך (הריבוע הכתום בצד ימין תחתון)



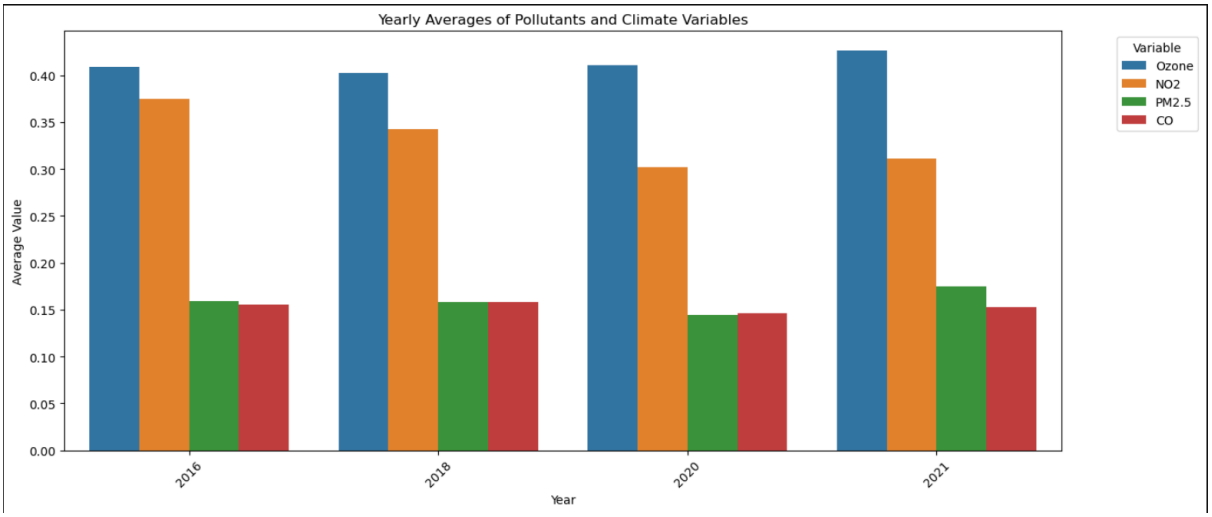
- גרף שבוחן קורלציה בין אוזון לטמפ' – ראינו בheatmap שיש קורלציה הכי חזקה בין אוזון למזג מבין כל המזהמים, לכן בחנו לבדוק את זה בגרף נפרד. בגרף ניתן לראות בברור שכאשר רמות האוזון עולות גם הטמפ מושפעת מכך ועולה גם כן



<u>מסקנה סופר חשובה – לפרט עליה בהרחבה!!</u>

- גרף BOXPLOT של המזהמים לפי שנים – בגרף זה כל קופסה הינו הIQR (טווח בין רבעוני) בין הרבע הראשון Q1 לרבע השלישי Q3 – ז"א זה הטווח בו נמצאים 50% מהנתונים, הזנבות זה המינימום והמקסימום שמחושבים ע"י:
  - $MIN = Q1 - 1.5 * IQR$
  - $MAX = Q3 + 1.5\, IQR$

  כל הערכים שנמצאים מחוץ לטווחים של המינימום והמקסימום (כל הנק בגרף) זה הערכים קיצוניים OUTLIERS. הקוו באמצע כל קופסה זה החציון (MEDIAN) – שמסומן גם בQ2. אפשר לראות שלמזהם 2.5PM יש הכי הרבה OUTLIERS ולאוזון הכי פחות, כמו כן אפשר לראות שבשנת 2020 החציון Q2 של רוב המזהמים הכי קטן. טוב להשערה שלנו (זה השנה של הסגר בקורונה)
- גרף עמודות שבודק את הערכים הממוצעים של כל מזהם לפי שנים. גם פה אפשר לראות שבשנת 2020 מרבית המזהמים היו הכי קטנים. גם טוב להשערה שלנו



	Year	Ozone	NO2	PM2.5	CO
0	2016	0.408613	0.374982	0.158722	0.155834
1	2018	0.402387	0.342724	0.158039	0.157934
2	2020	0.410736	0.302102	0.144801	0.146351
3	2021	0.425904	0.310940	0.174824	0.153056

רק עבור אוזון, בשנת 2020 כל שלושת המזהמים האחרים היו במינימום בשנת 2020.

Feature Engineering & Selection

עד עכשיו, רק נירמלנו את הפיצ'רים הנומריים שלנו (latitude & longitude חוץ מ) ל 0-1 ע"י MIN MAX SCALER.

השלב הבא של להעביר את הפיצ'רים שלא מספריים למספריים. כפי שאמרנו כבר את DATE, COUNTY העברנו למספרי ע"י ONE HOT ENCODING. חשוב לציין שלתאריך עשינו זאת עבור היום (ראשון יקבל 0 0 0 0 0 0 1), חודש (מרץ יקבל 0 0 0 0 0 0 0 0 0 1 0 0) ושנה (2016 יקבל 0 0 0 1).

אחר"כ השתמשנו בKMEANS כדי ליצור קלסטר לפיצ'רים של המיקום (latitude & longitude), קיבלנו 2 קלסטרים , לכן הערכים האפשרים לפיצ'ר זה גם בין 0-1 [או 0 או 1 תלוי לאן משוייך], ואגב יש שתי קלסטרי כיוון שבשלב הראשון נשארנו רק עם המחוזות Bronx, Queens.

בסיום השלב הזה הצלחנו להגדיל את כמות הפיצ'רים מ21 ל42 (3 מתוכם זה הLABELS – TEMP, PERCEP,RAIN) – סהכ הצלחנו להגדיל מ 18 ל 39. סהכ 2896 שורות בדאטסט הסופי.

(AND BENCHMARK WITH FAMOUS MODELS) ML Model For Weather Prediction

השלב האחרון זה לבנות את המודל הסופי – מודל מבוסס רשת ניורונים.

המודל נקרא VISL(LIDOR, SHAY, ILAY,VICTOR)

המטריקות שנבחן איתו את המודל הינם MAE, R^2,EVS .

כדי לבדוק כמה המודל שלנו יציב בתוצאות שלא, לא אימנו פעם אחת רק. אלה יצרנו לולאה שרצה 10 פעם, בכל איטרציה יוצרת אובייקט חדש של VISL, מאמנת אותו, ועושה EVALUATION, את התוצאות שמרנו בתוך מילון (ז"א מילון עם 3 מפתחות, אחד לכל מטריקה, והVLAUE של כל מפתח הינו רשימה בגודל 10 עם 10 תוצאות עבור אותה מטריקה).

אבל הביצועים של המודל השוונו לשתי מודלים קיימים RANDOM &LINEAR REGRESSION FOREST. גם להם יצרנו מילון דומה עם 10 אוולואציות.

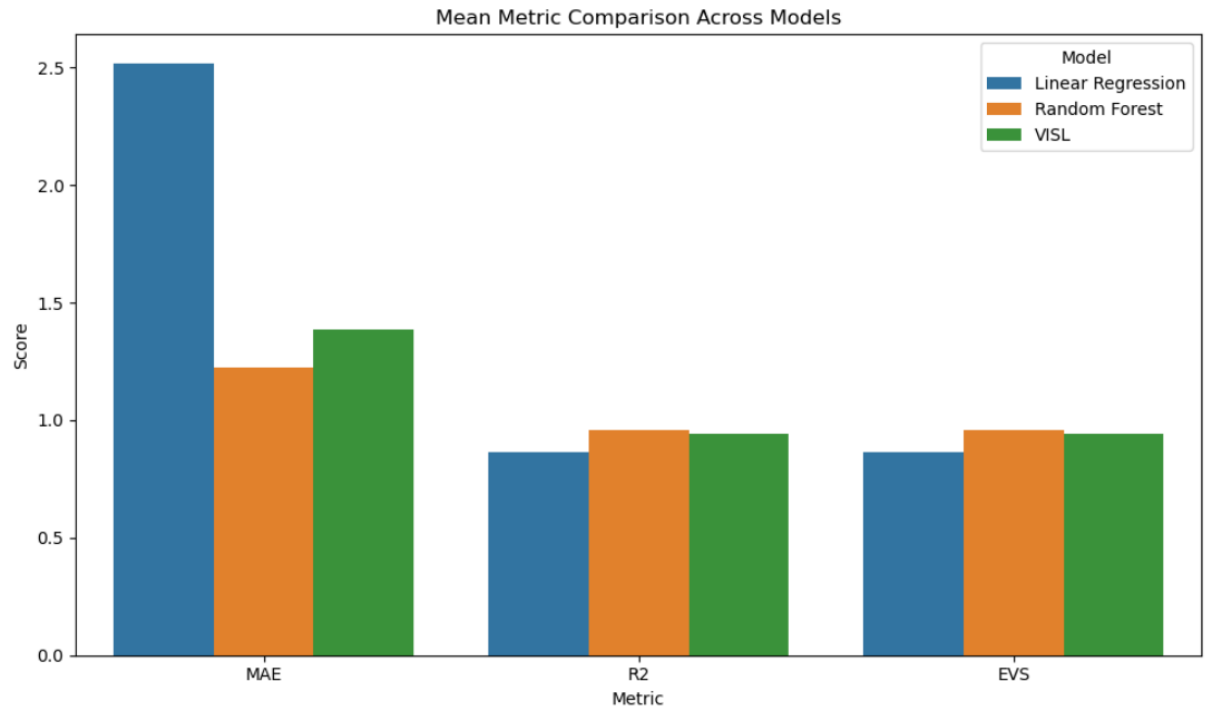
לבסוף השוונו בין הביצועים של המודלים בשתי שיטות:

- חישבנו מספר ערכים סטטיסטיים עבור כל מטריקה של מודל: ממוצע, שונות וסטיית תקן – והשוונו בין המודלים על סמך הערכים הללו (הערכים חושבו עבור אותן רשימות של תוצאות עבור כל מודל, למשל הערך של שונות עבור מודל VISL הינו פשוט שונות שחושבה לרשימה של 10 התוצאות עבור מודל VISL – יהיו 3 תוצאות, אחת לכל מטריקה – MAE, R^2, EVS)

```
STATISTICS FORMAT (Mean, Varicance, STD)
-----------------------------------------------------
Liner Regression Statistics:
{'mae': [2.5144, 0.0, 0.0], 'r2': [0.8642, 0.0, 0.0], 'evs': [0.8643, 0.0, 0.0]}

Random Forest Statistics:
{'mae': [1.22578, 0.00034, 0.01855], 'r2': [0.9555, 0.0, 0.00107], 'evs': [0.95565, 0.0, 0.0011]}

VISL Model Statistics:
{'mae': [1.384, 0.0078, 0.0886], 'r2': [0.9414, 0.0, 0.006], 'evs': [0.9433, 0.0, 0.0054]}
```



<u>דברים שהיינו ממליצים לעשות בהמשך (אל תעשו זה סתם לרשום):</u>

- מורידים את הפיצ'רים שקשורים למגז אוויר שראינו שיש להם קורלציה חזקה בינם לבין עצמם (multicollinearity) , מנסים לאסוף עוד פיצ'רים שקשורים למזהמים, ומאמנים מודל מחדש – יתכן שהעובדה שיש multicollinearity השפיעה על תוצאות המודל שלנו.
- מנסים לאסוף עוד פיצ'רים אחרים על NYC שמשתנים (למשל כמות המוניות הממוצעת שהיו במהלך היום, כמות העסקים שהיו פתוחים באותו יום, כמות מדד כל שהוא שבוחן את כמות האנשים שהיו בNYC ביום מסויים (בחוץ)) וכ'ו ...
- מנסים לשפר את המודל שלנו אולי עם מודל קיים מסויים שהיינו עושים לו FINE TUNE למשימה הספציפית שלנו