**מבוא למדע הנתונים פרויקט סיום קורס –**

**"ערי העולם וזיהום האוויר: תמונה גלובלית"**

מגישים :

שגיא מלכה – 203576053

שקד מזרחי - 322796269

אלינור לוי – 211563234

דניאל בן עמי – 313577140

**סקירת נתונים - updated\_global\_air\_pollution\_dataset**

**הקובץ מכיל נתונים על זיהום אוויר ממספר ערים ברחבי העולם. הנה כמה נקודות ותובנות עיקריות:**

- סה"כ רשומות: 23,463

- עמודות: 12, כולל:

- Country: המדינה בה נאספו הנתונים.

- City: העיר בה נאספו הנתונים.

- AQI Value: ערך מדד איכות האוויר הכללי.

- AQI Category: קטגוריית מדד איכות האוויר (לדוגמה, טוב, מתון).

- CO AQI Value: ערך מדד איכות האוויר של פחמן חד-חמצני.

- CO AQI Category: קטגוריית מדד הפחמן החד-חמצני.

- Ozone AQI Value: ערך מדד איכות האוויר של אוזון.

- Ozone AQI Category: קטגוריית מדד האוזון.

- NO2 AQI Value: ערך מדד איכות האוויר של תחמוצות חנקן.

- NO2 AQI Category: קטגוריית מדד תחמוצות החנקן.

- PM2.5 AQI Value: ערך מדד איכות האוויר של חלקיקים דקים (PM2.5).

- PM2.5 AQI Category: קטגוריית מדד חלקיקים דקים.

**תובנות עיקריות**

1. ערכים חסרים:

- ישנם מספר רב של ערכים חסרים בכמה עמודות. לדוגמה:

- Country חסר ב-5,560 רשומות.

- AQI Value חסר ב-5,461 רשומות.

- CO AQI Value חסר ב-5,488 רשומות.

- עמודות מסוימות, כמו Ozone AQI Value ו-PM2.5 AQI Value, חסרות פחות ערכים אך עדיין מספיק כדי להשפיע על הניתוח.

2. סוגי נתונים:

- רוב העמודות הן מסוג object, מה שמצביע שהן כוללות מחרוזות או סוגים מעורבים.

- שתי עמודות, Ozone AQI Value ו-PM2.5 AQI Value, הן מסוג float64.

3. בעיות אפשריות באיכות הנתונים:

- ישנם ערכים לא מספריים בעמודות AQI Value ו-NO2 AQI Value, שיש לטפל בהם במהלך ניקוי הנתונים.

- יש ערכים כמו 9999 בעמודת Country, שיש לתקן או להסיר.

**דוגמאות לנתונים**

להלן מספר שורות ראשונות מהקובץ כדי לתת תחושה של מבנה הנתונים:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Country | City | AQI Value | AQI Category | CO AQI Value | CO AQI Category | Ozone AQI Value | Ozone AQI Category | NO2 AQI Value | NO2 AQI Category | PM2.5 AQI Value | PM2.5 AQI Category |
| NaN | Praskoveya | 51 | Moderate | 1 | Good | NaN | Good | 0 | Good | 51.0 | NaN |
| 9999 | Presidente Dutra | NaN | Good | NaN | Good | 5.0 | Good | 1 | Good | NaN | NaN |
| 9999 | Priolo Gargallo | NaN | NaN | 1 | Good | 39.0 | NaN | 2 | Good | 66.0 | Moderate |
| Poland | Przasnysz | xsads | NaN | NaN | 9999 | 34.0 | Good | NaN | NaN | 20.0 | Good |
| NaN | 9999 | 22 | Good | 0 | NaN | 22.0 | Good | 0 | NaN |  |  |

**כדי להפוך את הקובץ ליותר שימושי לניתוח, יש לשקול את השלבים הבאים:**

1. ניקוי נתונים:

- לטפל בערכים חסרים באמצעות מילוי בערכים סטטיסטיים מתאימים (ממוצע, חציון) או הסרת שורות/עמודות עם נתונים חסרים רבים.

- להמיר ערכים לא מספריים למספריים ולהבטיח טיפול נכון בערכים לא תקינים.

- להחליף ערכים כמו 9999 ב-NaN או בתיקונים מתאימים.

2. שינוי נתונים:

- לתקן את ערכי ה-AQI כדי להבטיח עקביות.

- לבצע קיבוץ נוסף של הנתונים אם נדרש, כמו קיבוץ ערים לפי אזורים או מדינות.

3. ניתוח נתונים:

- לבצע ניתוח נתונים חוקר (EDA) כדי לגלות דפוסים, מגמות וקורלציות.

- להשתמש בטכניקות ויזואליזציה כדי להבין טוב יותר את ההתפלגות של ערכי ה-AQI בערים ובמדינות השונות.

**סקירת נתונים - worldcities**

**הקובץ מכיל נתונים על ערים ברחבי העולם. הנה כמה נקודות ותובנות עיקריות:**

- סה"כ רשומות: 15,493

- עמודות: 11, כולל:

- City: שם העיר.

- City\_ascii: שם העיר באותיות לטיניות.

- Lat: קו רוחב של העיר.

- Lng: קו אורך של העיר.

- Country: המדינה בה נמצאת העיר.

- iso2: קוד המדינה בן שתי אותיות.

- iso3: קוד המדינה בן שלוש אותיות.

- admin\_name: שם המחוז או האזור המנהלי.

- capital: האם העיר היא עיר בירה.

- population: אוכלוסיית העיר.

- id: מזהה ייחודי לעיר.

**תובנות עיקריות**

1. ערכים חסרים:

- ישנם מספר ערכים חסרים בכמה עמודות. לדוגמה:

- City\_ascii חסר ב-1,549 רשומות.

- Lat חסר ב-1,549 רשומות.

- Country חסר ב-1,548 רשומות.

- population חסר ב-3,081 רשומות.

- עמודת capital כוללת הרבה ערכים חסרים (רק 4,709 רשומות מכילות ערכים).

2. סוגי נתונים:

- רוב העמודות הן מסוג object, כלומר מחרוזות.

- שתי עמודות, Lat ו-id, הן מסוג float64.

3. בעיות אפשריות באיכות הנתונים:

- בעמודת population, יש ערכים לא מספריים כגון xxx, שיש לטפל בהם במהלך ניקוי הנתונים.

- ערכי Lng (קו אורך) הם מחרוזות במקום מספרים, ויש להמירם לסוג הנתונים המתאים.

**דוגמאות לנתונים**

להלן מספר שורות ראשונות מהקובץ כדי לתת תחושה של מבנה הנתונים:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| City | City\_ascii | Lat | Lng | Country | iso2 | iso3 | admin\_name | capital | population | id |
| Tokyo | Tokyo | 35.6850 | 139.7514 | Japan | JP | JPN | Tōkyō | primary | 35676000 | 1.392686e+09 |
| New York | NaN | 40.6943 | -73.9249 | United States | NaN | USA | New York | NaN | xxx | 1.840034e+09 |
| Mexico City | NaN | 19.4424 | -99.131 | Mexico | NaN | MEX | Ciudad de México | primary | 19028000 | 1.484248e+09 |
| Mumbai | Mumbai | 19.0170 | 72.857 | India | IN | IND | NaN | admin | 18978000 | 1.356227e+09 |
| São Paulo | Sao Paulo | -23.5587 | -46.625 | 122 | BR | BRA | NaN | admin | 18845000 | 1 |

**כדי להפוך את הקובץ ליותר שימושי לניתוח, יש לשקול את השלבים הבאים:**

1. ניקוי נתונים:

- לטפל בערכים חסרים באמצעות מילוי בערכים סטטיסטיים מתאימים (ממוצע, חציון) או הסרת שורות/עמודות עם נתונים חסרים רבים.

- להמיר ערכים לא מספריים למספריים ולהבטיח טיפול נכון בערכים לא תקינים.

- להמיר את עמודת Lng למספרים (float).

2. שינוי נתונים:

- לבצע קיבוץ נוסף של הנתונים אם נדרש, כמו קיבוץ ערים לפי אזורים או מדינות.

3. ניתוח נתונים:

- לבצע ניתוח נתונים חוקר (EDA) כדי לגלות דפוסים, מגמות וקורלציות.

- להשתמש בטכניקות ויזואליזציה כדי להבין טוב יותר את ההתפלגות של אוכלוסיות בערים השונות ובמדינות השונות.

גרף מס 1 :

גרף עמודות של ממוצע ערכי AQI לפי מדינה

הגרף מציג את ממוצע ערכי ה-AQI (אינדקס איכות אוויר) לפי מדינה. הנה כמה דפוסים ומסקנות שניתן להפיק מהגרף:

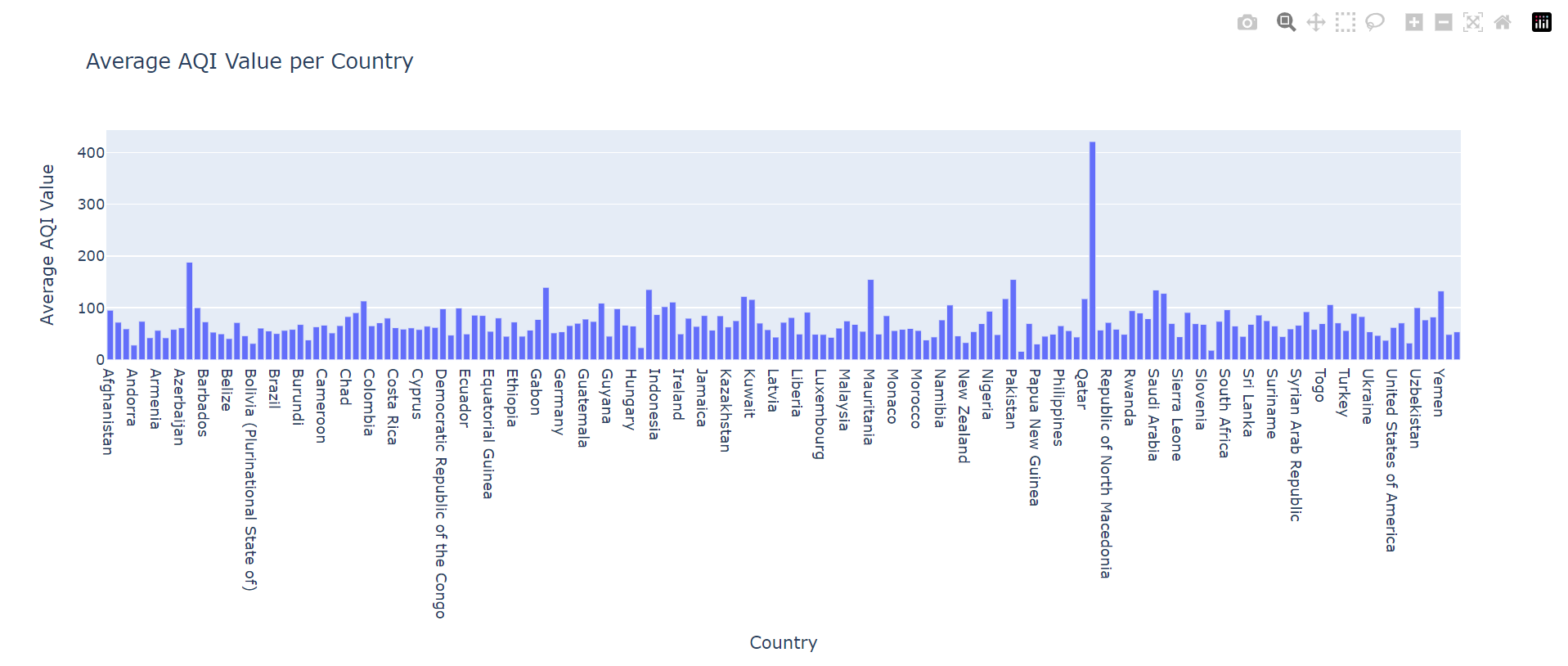
ערכי AQI גבוהים במדינות מסוימות:

ניתן לראות שישנן מספר מדינות עם ערכי AQI גבוהים במיוחד (מעל 200), כמו קטר והפיליפינים.

מדינות אלו עלולות להתמודד עם בעיות זיהום אוויר חמורות שיכולות להשפיע על בריאות התושבים.

מדינות עם ערכי AQI נמוכים:

רוב המדינות מציגות ערכי AQI נמוכים יחסית (פחות מ-100).

מדינות אלו נהנות מאיכות אוויר טובה יחסית

גרף מס 2:

קורלציה בין PM2.5 AQI Value ל-AQI Value:

ישנה קורלציה גבוהה (0.75) בין ערכי PM2.5 AQI לערכי AQI כלליים. זה מצביע על כך ש-PM2.5 תורם משמעותית לערכי ה-AQI הכלליים בערים.

קורלציה בין CO AQI Value ל-AQI Value:

ישנה קורלציה בינונית (0.34) בין ערכי CO AQI לערכי AQI כלליים. CO הוא גם גורם תורם לערכי ה-AQI הכלליים, אך לא בצורה משמעותית כמו PM2.5.

קורלציה בין ערכי AQI שונים:

ערכי PM2.5 AQI ו-CO AQI הם בעלי קורלציה גבוהה ביניהם (0.34).

ערכי NO2 AQI ו-CO AQI הם בעלי קורלציה בינונית (0.37).

ערכי Ozone AQI הם בעלי קורלציה נמוכה עם שאר המזהמים, למעט CO AQI (0.12).

קורלציה בין אוכלוסייה למזהמים:

יש קורלציה נמוכה מאוד בין גודל האוכלוסייה לערכי ה-AQI והמזהמים השונים. ערכים אלה הם קרובים ל-0, מה שמצביע על כך שגודל האוכלוסייה אינו קשור ישירות לערכי הזיהום הנמדדים.

ערכים עצמיים:

כל ערך של קורלציה עם עצמו הוא 1, מה שמצביע על קורלציה מושלמת.

מסקנות כלליות

השפעת PM2.5 על AQI הכללי:

PM2.5 הוא אחד הגורמים המשמעותיים ביותר המשפיעים על ערכי ה-AQI הכלליים.

גורמים נוספים המשפיעים על AQI:

CO הוא גם גורם תורם משמעותי לערכי ה-AQI הכלליים, אך השפעתו נמוכה מזו של PM2.5.

השפעה מועטה של אוכלוסייה על זיהום:

גודל האוכלוסייה לא משפיע באופן ישיר על רמות הזיהום הנמדדות.

קורלציות בין מזהמים:

קשרים בין סוגי מזהמים שונים יכולים לסייע בזיהוי מקורות זיהום משותפים ודרכים לטיפול בזיהוםתמונה שמכילה צילום מסך, טקסט, ריבוע, מלבן

התיאור נוצר באופן אוטומטי

גרף מס 3:

ניתוח ה-Violin Plot של ערכי AQI לפי מדינה

גרף ה-Violin Plot מציג את התפלגות ערכי AQI עבור כל מדינה. זהו גרף שמשלב היסטוגרמה ו-box plot, והוא מאפשר לראות את צפיפות הנתונים ואת ההתפלגות הכללית שלהם בצורה ויזואלית.

מסקנות ודפוסים מהמטריצה:

התפלגות רחבה של ערכי AQI במדינות רבות:

ניתן לראות שבמדינות רבות יש טווח רחב של ערכי AQI. לדוגמה, ישנן מדינות שבהן ערכי ה-AQI מגיעים עד ל-500.

התפלגות רחבה זו מצביעה על כך שישנם אזורים בתוך המדינות הללו עם רמות זיהום שונות מאוד.

ערכי AQI קיצוניים:

ישנן מספר מדינות עם ערכים קיצוניים של AQI. הערכים הגבוהים הללו מצביעים על בעיות זיהום חמורות באזורים מסוימים בתוך המדינות הללו.

ערכים נמוכים של AQI:

ניתן לראות שגם במדינות עם ערכים גבוהים מאוד, יש אזורים עם ערכי AQI נמוכים מאוד.

זה מצביע על כך שישנם אזורים בתוך המדינות הללו שמצליחים לשמור על איכות אוויר טובה יחסית.

פיזור הערכים בין המדינות:

תמונה שמכילה טקסט, צילום מסך, קו, עלילה

התיאור נוצר באופן אוטומטיהפיזור הרחב של ערכי AQI בין המדינות מצביע על כך שיש הבדלים משמעותיים ברמת הזיהום בין המדינות השונות.

גרף מס 4:

ניתוח גרף פיזור צבעוני של גודל אוכלוסייה מול ערכי זיהום אוויר AQI

גרף פיזור זה מציג את הקשר בין גודל האוכלוסייה לערכי ה-AQI (אינדקס איכות האוויר) בערים שונות. הצבעים של הנקודות בגרף מייצגים את ערכי ה-AQI, כאשר צבעים כהים יותר מציינים ערכים גבוהים יותר

מסקנות כלליות

תמונה שמכילה טקסט, תרשים, עלילה, צילום מסך

התיאור נוצר באופן אוטומטיהגרף מציג תמונה מורכבת של הקשר בין גודל האוכלוסייה לערכי ה-AQI בערים שונות. נראה שאין קורלציה ישירה בין גודל האוכלוסייה לערכי ה-AQI, וישנם ערכים גבוהים מאוד של AQI בערים עם אוכלוסיות קטנות יחסית. ניתוח מעמיק יותר יכול לסייע בהבנת הגורמים והשלכות הזיהום, ולקדם צעדים לשיפור איכות האוויר בערים שונות.

גרף מס 5:

ניתוח גרף עמודות של ערכי AQI עבור 20 הערים הגדולות ביותר

גרף זה מציג את ערכי ה-AQI (אינדקס איכות האוויר) עבור 20 הערים הגדולות ביותר לפי גודל האוכלוסייה.

מסקנות ודפוסים מהגרף:

ערכים גבוהים במיוחד של AQI:

סיאול (Seoul) ופריז (Paris) מציגות ערכי AQI גבוהים מאוד, מעל 400.

ערכים אלו מצביעים על בעיות זיהום אוויר חמורות באזורים אלו.

ערכים בינוניים של AQI:

ערים כמו שנחאי (Shanghai), סאו פאולו (São Paulo) ובייג'ינג (Beijing) מציגות ערכי AQI בינוניים (בין 100 ל-200).

זה מצביע על רמות זיהום אוויר מסוימות, אך לא קיצוניות כמו סיאול ופריז.

ערכים נמוכים של AQI:

ערים כמו טוקיו (Tokyo), מקסיקו סיטי (Mexico City) ומומבאי (Mumbai) מציגות ערכי AQI נמוכים יחסית (פחות מ-100). זה מצביע על איכות אוויר טובה יחסית בערים אלו.

פיזור ערכי ה-AQI בין הערים:

הפיזור הרחב של ערכי AQI בין הערים מצביע על כך שיש הבדלים משמעותיים ברמות הזיהום בין הערים השונות.

תמונה שמכילה טקסט, צילום מסך, קו, תרשים

התיאור נוצר באופן אוטומטיערכים אלו יכולים לנבוע מגורמים שונים כמו תחבורה, תעשייה, ומדיניות סביבתית

גרף מס 6:

ניתוח גרף עמודות של גודל האוכלוסייה עבור 20 הערים הגדולות ביותר (ללא ערכים לא ידועים)

גרף זה מציג את גודל האוכלוסייה של 20 הערים הגדולות ביותר בעולם, תוך התעלמות מערכים לא ידועים.

מסקנות ודפוסים מהגרף:

טוקיו בראש הרשימה:

טוקיו היא העיר הגדולה ביותר מבחינת גודל האוכלוסייה, עם אוכלוסייה של מעל 35 מיליון תושבים.

זה מצביע על כך שטוקיו היא מטרופולין גדול ומשמעותי מאוד מבחינה כלכלית, תרבותית וחברתית.

ערים עם אוכלוסייה גדולה מעל 10 מיליון תושבים:

מקסיקו סיטי, מומבאי, סאו פאולו ושנחאי הן גם ערים עם אוכלוסיות גדולות מאוד, כל אחת עם יותר מ-15 מיליון תושבים.

ערים אלו הן מרכזים כלכליים חשובים ומרכזים של פעילות מסחרית ותעשייתית.

ערים עם אוכלוסייה בינונית בגודל:

ערים כמו לוס אנג'לס, קרצ'י, ריו דה ז'ניירו ובייג'ינג מציגות אוכלוסיות בינוניות בגודלן, בין 10 ל-15 מיליון תושבים.

ערים אלו הן גם מרכזים חשובים אך פחות עמוסים מבחינת גודל אוכלוסייה בהשוואה לטוקיו ומקסיקו סיטי.

ערים קטנות יחסית מבין הגדולות ביותר:

ערים כמו ג'קרטה, לונדון, גואנגג'ואו ושיקגו נמצאות בחלק התחתון של הרשימה מבחינת גודל האוכלוסייה, כל אחת עם פחות מ-10 מיליון תושבים.

תמונה שמכילה טקסט, צילום מסך, קו, תרשים

התיאור נוצר באופן אוטומטילמרות שהן קטנות יחסית לערים הגדולות ביותר, הן עדיין מרכזים חשובים כלכליים ותרבותיים

גרף מס 7:

ניתוח היסטוגרמה של התפלגות ערכי AQI

ההיסטוגרמה מציגה את התפלגות ערכי ה-AQI (אינדקס איכות האוויר) בערים שונות.

מסקנות ודפוסים מהגרף:

ערכי AQI נפוצים:

רוב ערכי ה-AQI מרוכזים בטווח של 0 עד 100.

זה מצביע על כך שבמרבית הערים יש איכות אוויר טובה יחסית.

ערכים גבוהים של AQI:

ישנה ירידה חדה במספר הערכים מעל 100.

ערכים גבוהים מ-100 הם נדירים יותר, ומצביעים על זיהום אוויר חמור יותר.

פיזור ערכי ה-AQI:

ניתן לראות שהתפלגות ערכי ה-AQI נוטה לכיוון הערכים הנמוכים, עם מספר מועט של ערכים גבוהים מאוד (מעל 300).

זה מצביע על כך שרוב הערים מצליחות לשמור על רמות זיהום נמוכות יחסית.

זיהום אוויר קיצוני:

ישנם מספר ערכים קיצוניים (מעל 400), אך הם נדירים מאוד.

תמונה שמכילה טקסט, צילום מסך, עלילה, קו

התיאור נוצר באופן אוטומטיערכים אלו מצביעים על מצבי זיהום אוויר חריגים ביותר, שיכולים להיות תוצאה של אירועים ספציפיים כמו תעשייה כבדה, שריפות, או תנאי מזג אוויר קיצוניים

גרף מס 8:

ניתוח גרף מגמות ערכי AQI לערים נבחרות

הגרף מציג את המגמות של ערכי ה-AQI (אינדקס איכות האוויר) עבור ארבע ערים נבחרות: פריז, ניו יורק, טוקיו ומומבאי.

מסקנות ודפוסים מהגרף:

ערכים יציבים יחסית:

ערכי ה-AQI של כל ארבע הערים נראים יציבים יחסית לאורך הזמן. אין שינויים חדים או מגמות ברורות בערכים המוצגים בגרף.

כל הערכים נמצאים בטווח צר מאוד (בין 71 ל-72), מה שמצביע על יציבות ברמות זיהום האוויר בערים הללו בתקופה המדוברת.

הבדלים מינימליים בין הערים:

אין הבדלים משמעותיים בערכי ה-AQI בין ארבע הערים הנבחרות.

תמונה שמכילה טקסט, קו, תרשים, צילום מסך

התיאור נוצר באופן אוטומטיכל הערים מציגות ערכים דומים מאוד, מה שמצביע על כך שאיכות האוויר בערים אלו היא דומה בתקופה המדוברת

גרף מס 9:

ניתוח Pairplot של תכונות נבחרות

הגרף מציג Pairplot של מספר תכונות נבחרות, כולל גודל האוכלוסייה, ערכי ה-AQI הכלליים, וערכי AQI של מזהמים שונים כמו PM2.5, אוזון, NO2 ו-CO. Pairplot זה מאפשר לנו לראות את הקשרים בין כל זוג משתנים.

מסקנות ודפוסים מהגרף:

קורלציה בין ערכי AQI שונים:

ניתן לראות קורלציה ברורה בין ערכי AQI כלליים לערכי PM2.5 AQI (כפי שניתן לראות במטריצת הקורלציה שהוצגה קודם לכן).

גם בין ערכי CO AQI לערכי AQI כלליים ניתן לראות קשר מסוים, אם כי פחות חזק.

פיזור ערכי AQI לפי גודל האוכלוסייה:

ניתן לראות שאין קשר ברור בין גודל האוכלוסייה לערכי ה-AQI. הפיזור של הערכים רחב, ומכאן ניתן להסיק שגודל האוכלוסייה אינו הגורם העיקרי המשפיע על ערכי ה-AQI.

זה מתיישב עם מה שראינו קודם בגרף הפיזור הצבעוני.

קורלציה בין מזהמים שונים:

ניתן לראות קשרים מסוימים בין מזהמים שונים. לדוגמה, יש קשר מסוים בין PM2.5 ו-CO, ו-NO2 ו-CO.

קשרים אלו מצביעים על כך שמקורות זיהום מסוימים פולטים יותר מסוג אחד של מזהם בו-זמנית.

קשרים בין מזהמים לערכי AQI כלליים:

PM2.5 הוא מזהם עם הקשר החזק ביותר לערכי ה-AQI הכלליים. זה משקף את ההשפעה הגדולה של חלקיקים אלה על איכות האוויר הכללית.

תמונה שמכילה טקסט, מפה, תרשים, קו

התיאור נוצר באופן אוטומטיגם CO תורם במידה מסוימת לערכי ה-AQI הכלליים

גרף מס 10:

ניתוח גרף עוגה של התפלגות האוכלוסייה לפי מדינה

גרף העוגה מציג את ההתפלגות של האוכלוסייה בין המדינות השונות. המדינות עם האוכלוסייה הגדולה ביותר מסומנות בצבעים שונים.

מסקנות ודפוסים מהגרף:

אוכלוסיות גדולות במיוחד:

הודו (India) וסין (China) הן המדינות עם האוכלוסייה הגדולה ביותר, עם חלקים גדולים יחסית בעוגה.

זה משקף את העובדה שהודו וסין הן המדינות המאוכלסות ביותר בעולם.

ארצות הברית:

ארצות הברית (United States of America) גם היא בעלת חלק נכבד באוכלוסייה הגלובלית, אם כי קטן יותר מזה של הודו וסין.

מדינות עם אוכלוסייה בינונית וקטנה:

רוב המדינות האחרות תורמות חלק קטן יחסית להתפלגות האוכלוסייה העולמית.

ניתן לראות שיש פיזור רחב של מדינות עם חלקים קטנים מאוד, דבר שמצביע על כך שיש מדינות רבות עם אוכלוסיות קטנות יותר

תמונה שמכילה טקסט, צילום מסך, עיצוב

התיאור נוצר באופן אוטומטי

**ניתוח גרף Elbow Method וגרף Clusters**

גרף Elbow Method למציאת מספר האשכולות האופטימלי

גרף זה מציג את ה-Sum of Squared Distances (SSE) עבור מספר אשכולות (k) שונה. נקודת הברך בגרף היא המקום שבו הירידה ב-SSE מתחילה להיות פחות תלולה, ומספר האשכולות האופטימלי נבחר בנקודה זו.

מסקנות מגרף Elbow Method:

ניתן לראות שנקודת הברך בגרף היא בסביבות 4 אשכולות. זהו מספר האשכולות שבו הירידה ב-SSE מתחילה להיות פחות תלולה.

מספר האשכולות האופטימלי עבור הנתונים הוא 4.

גרף Clusters של אוכלוסייה וערכי AQI

גרף זה מציג את האשכולות של הנתונים המבוססים על אוכלוסייה וערכי AQI. כל נקודה מייצגת עיר, וצבעה מציין את האשכול שאליו היא שייכת.

מסקנות מגרף Clusters:

ניתן לראות את החלוקה לאשכולות שונים לפי צבעים.

רוב הנקודות מרוכזות בצד השמאלי של הגרף, בערים עם אוכלוסיות קטנות יחסית וערכי AQI משתנים.

ישנן נקודות בערכים גבוהים יותר של אוכלוסייה (מעל 1 מיליון) עם ערכי AQI נמוכים יחסית, מה שמצביע על איכות אוויר טובה בערים עם אוכלוסייה גדולה.

דפוסים שניתן לזהות:

אשכולות עיקריים:

אשכול אחד מרכז ערים עם אוכלוסייה קטנה וערכי AQI נמוכים.

אשכול נוסף מרכז ערים עם אוכלוסייה קטנה וערכי AQI גבוהים.

ישנו אשכול של ערים עם אוכלוסייה בינונית (עד 1 מיליון) וערכי AQI בינוניים.

ערכים קיצוניים:

ניתן לראות מספר ערים עם ערכי AQI גבוהים מאוד (מעל 400), בעיקר בערים עם אוכלוסיות קטנות.

הערים עם האוכלוסיות הגדולות ביותר מציגות ערכי AQI נמוכים יחסית, מה שמצביע על מדיניות אפקטיבית לניהול איכות האוויר בערים אלו

**ניתוח רגרסיה לינארית: אוכלוסייה מול ערכי AQI**

הגרף מציג רגרסיה לינארית בין גודל האוכלוסייה (ציר ה-X) לערכי ה-AQI (ציר ה-Y). קו הרגרסיה האדום מראה את הקשר הלינארי המשוער בין שני המשתנים.

מסקנות ודפוסים מהגרף:

קו רגרסיה עולה:

קו הרגרסיה האדום מראה מגמה עולה בין גודל האוכלוסייה לערכי ה-AQI.

המשמעות היא שככל שגודל האוכלוסייה עולה, ערכי ה-AQI נוטים לעלות גם כן.

פיזור הנתונים:

ניתן לראות פיזור רחב של הנתונים סביב קו הרגרסיה.

הפיזור הרחב מצביע על כך שהקשר בין גודל האוכלוסייה לערכי ה-AQI הוא לא חזק מאוד, וישנם גורמים נוספים שמשפיעים על ערכי ה-AQI.

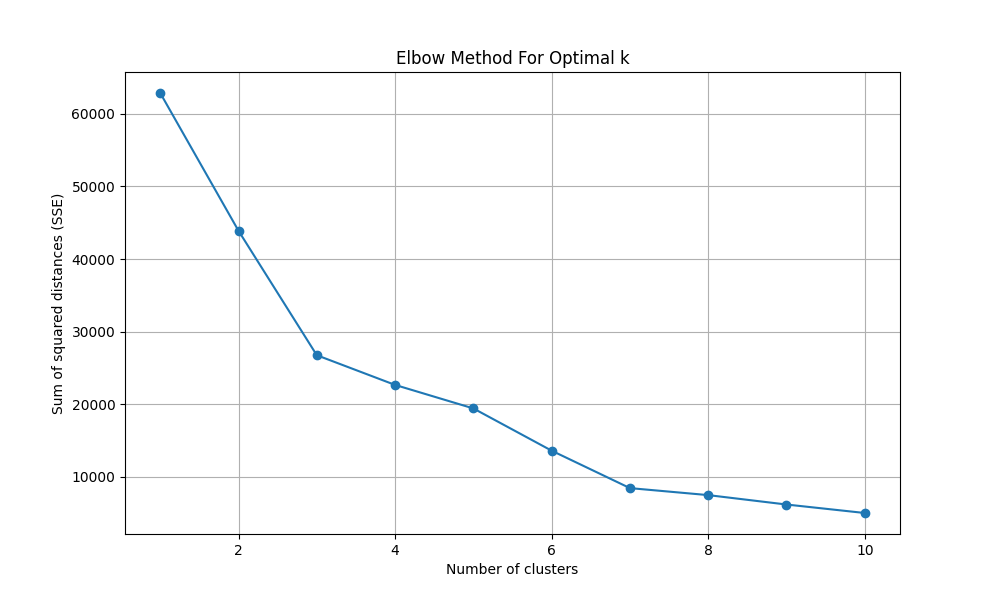
ערכי AQI גבוהים מאוד בערים קטנות:

ישנן נקודות עם ערכי AQI גבוהים מאוד (מעל 400) בערים עם אוכלוסיות קטנות יחסית.

זה מצביע על כך שישנם גורמים מקומיים שיכולים לגרום לערכי AQI גבוהים בערים קטנות.

ערכי AQI נמוכים בערים גדולות:

בערים עם אוכלוסיות גדולות (מעל 1 מיליון), ניתן לראות ערכי AQI נמוכים יחסית (פחות מ-200).

ייתכן שערים גדולות משקיעות יותר במערכות להפחתת זיהום האוויר ושיפור איכות האוויר

