**מחסני נתונים תרגיל 1 - COVID 19 בדרום קוריאה**

**חלק א- תכנון מחסן הנתונים**

**הקדמה**

KCDC (מרכזי קוריאה לבקרת מחלות ומניעה) פרסמה המידע לגבי התפשטות COVID-19 במהירות ברחבי קוריאה. התבקשנו לבנות מחסן נתונים מובנה המבוסס על חומרי הדו"ח של KCDC והממשלות המקומיות. על מנת לספק מידע שיעזור להתמודד עם התפשטות המחלה ברחבי קוריאה.

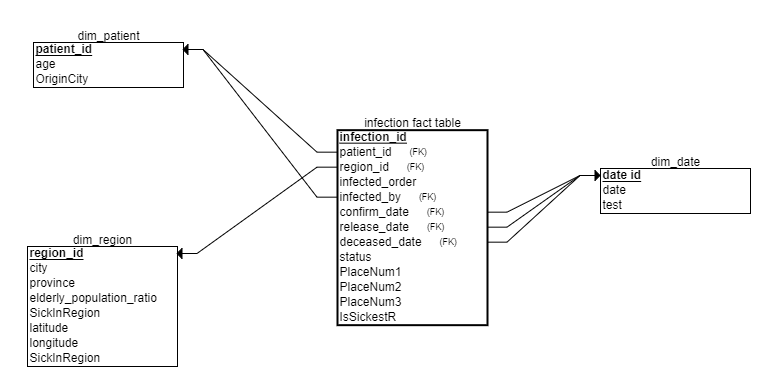
**שאילתות נדרשות-**

1. מספר החולים ביום בשלוש הערים עם שיעור ההדבקה הגבוה ביותר בדרום קוריאה.
2. מספר החולים המחלימים ביום בשלוש הערים עם שיעור ההדבקות הגבוה ביותר בדרום קוריאה.
3. מספר הערים בהן המטופל ביקר עד שלושה ימים, עד יומיים ויום לפני שנכנס לבידוד.
4. כמות האנשים בסיאול מתחת לגיל 30 שנדבקו וגם הדביקו באדם מעל גיל 30
5. צור מידע שיעזור לך להחליט אילו ערים בדרום קוריאה צריכות להיסגר כדי להפחית את התפשטות הנגיף. **המדיניות שלנו:**  
   סכום החולים הפעילים בכל אזור(עיר, מחוז) שאושרו מהתאריך 2020-03-05 והילך (טווח הזמן שמתחילים התסמינים) שיש בו יותר מ 20 חולים.(השתמשנו בשאילתה ובנוסף בוויזואליזציה).

מאפייני מחסן הנתונים:

1. **מחסן הנתונים שלנו מתמקד בהתקדמות המחלה בדרום קוריאה ע"י הסתכלות על ההדבקות(גרעין).  
   טבלת העובדות שלנו(fact) מכילה את ההדבקות הנ"ל. כל רשומה בטבלה(שורה) מייצגת הדבקה של אדם בקוריאה:**
   1. **מצבו הרפואי: בריא, מבודד(חולה),קורבן קורונה(נפטר)**
   2. **תאריכים קריטיים בהם אירע שינוי במצבו(או יתרחש שינוי עתידי): תאריך אישור מחלה, תאריך החלמה, תאריך פטירה.**
   3. **מידע בנוגע להדבקה: מאיזה אדם נדבק, סדר ההדבקה שלו**
   4. **מידע בנוגע להדבקות שיכל לבצע בשלושת הימים לפי אישור המחלה: מספר הערים בהן ביקר אותו אדם עבור שלושת הימים.**
2. **המימדים אשר בחרנו להכיל במחסן הנתונים שלנו:**
   1. Date **– מימד תאריך**
   2. Patient **– מימד אדם(פציינט)**
   3. Region **-מימד אזור(תת מחוז) , בחרנו להשתמש בשילוב של העיר והמחוז כיוון ששמנו לב כי העמודה שקיבלנו מה**CSV **בשם** Code **הכילה את הקוד המנהלי הייחודי של כל תת מחוז/עיר ולכן תייצג בצורה מיטבית את קושי בהגדרה של "עיר" ולא תהרוס את הנתונים.**
3. **נדגיש כי קיימות עמודות אשר השתמשנו בהן פעמים מעטות כיוון שרצינו להשאיר מידע שלדעתנו ישמש עבור שאילתות עתידיות ויאפשר למחסן הנתונים גמישות עתידית(יפורט בהנחות לאחר טבלאות הפירוט)**

**סכמת כוכב של מחסן הנתונים:**

****

**טבלאות מימדים ועובדות**

|  |  |  |  |
| --- | --- | --- | --- |
| **טבלת מימד(patient)** | | | |
| **שדה** | **טיפוס נתונים** | **מקור נתונים** | **הערות** |
| patient\_id | BIGINT | דה מקובץ Patient | מפתח ראשי ייחודי |
| Age | int | מחושב על ידי חיסור תאריך הנוכחי מbirth\_year מקובץ Patient | תאריך הנתונים |
| OriginCity | Varchar | שלף מקובץ Region מעמודה City | העיר שבה גר החולה/מחלים |

|  |  |  |  |
| --- | --- | --- | --- |
| **טבלת מימד(region)** | | | |
| **שדה** | **טיפוס נתונים** | **מקור נתונים** | **הערות** |
| region\_id | BIGINT | שדה Auto Increment | מפתח ראשי, ייחודי, מזהה. הקוד המנהלי של התת-מחוז/עיר. |
| City | Varchar | נשלף מקובץ Region מעמודה City | העיר המרכזית בRegion |
| Province | Varchar | נשלף מקובץ Region מעמודה Province |  |
| elderly\_population\_ratio | Varchar | נשלף מקובץ Region מעמודה elderly\_population\_ratio | אחוז האוכלוסייה המבוגרת במחוז |
| SickInRegion | int | מחושב ע"י סכימת עמודה "State" מקובץ Pantient כאשר העמודה שווה ל"isolated" ולפי העמודה City | מספר ה"חולים" יומי לפי מחוז |
| latitude | real | נשלף מקובץ Region מעמודה latitude | קו אורך על פני כאורדינטות כדור הארץ. שימשו עיקרי ב**וויזואליזציה**. |
| longitude | real | נשלף מקובץ Region מעמודה longitude | קו רוחב על פני כאורדינטות כדור הארץ. שימשו עיקרי ב**וויזואליזציה**. |

|  |  |  |  |
| --- | --- | --- | --- |
| **טבלת מימד(Date)** | | | |
| **שדה** | **טיפוס נתונים** | **מקור נתונים** | **הערות** |
| Date\_id | int | שדה Auto Increment | מפתח ראשי, ייחודי, מזהה |
| Date | Date | נשלף מקובץ time עמודה Date . מתחיל מהתאריך הכי קטן ועד למקסימלי | תאריך הנתונים |
| test | int | נשלף מקובץ time עמודה test. | מספר הבדיקות ליום. \*שימוש עתידי למעקב ולגרפים אשר יראו האם מתבצעות מספיק בדיקות במאבק במחלה. |

|  |  |  |  |
| --- | --- | --- | --- |
| **טבלת עובדות(infection)** | | | |
| **שדה** | **טיפוס נתונים** | **מקור נתונים** | **הערות** |
| infection\_id | int | שדה Auto Increment | מפתח ראשי, ייחודי, מזהה |
| patient\_id | BIGINT | דה מקובץ Patient | מפתח זר ממימד(patient) |
| Region\_id | Varchar | שדה מקובץ Patient | מפתח זר ממימד(Region) |
| infected\_order | int | שדה מקובץ Patient | סדר ההדבקה בשרשרת |
| infected\_by | BIGINT | שדה מקובץ Patient | parient\_idמפתח זר של  מי שהדביק את הנ”ל |
| confirm\_date | INT | מפתח ראשי מימד Date | מפתח זר ממימד(Date) |
| release\_date | INT | מפתח ראשי מימד Date | מפתח זר ממימד(Date) |
| deceased\_date | INT | מפתח ראשי מימד Date | מפתח זר ממימד(Date) |
| status | int | מחושב על פי עמודה "State" | מקרא: 0- משוחרר  1- חולה אפשרי(בבידוד)  2- נפטר |
| PlaceNum1 | int | מחושב ע"י סכימת המקומות בהם היה הpatient ביום לפני ההידבקות מקובץ route | מספר הערים/מחוזות בהם ביקר הפציינט כיום לפני הדבקתו |
| PlaceNum2 | int | מחושב ע"י סכימת המקומות בהם היה הpatient יומיים לפני ההידבקות מקובץ route | מספר הערים/מחוזות בהם ביקר הפציינט כיומיים לפני הדבקתו |
| PlaceNum3 | int | מחושב ע"י סכימת המקומות בהם היה הpatient שלושה ימים לפני ההידבקות מקובץ route ע"י סכימת כמות הCity הזהים. | מספר הערים/מחוזות בהם ביקר הפציינט כשלושה ימים לפני הדבקתו. |
| IsSickestR | int | מחושב ע"י בדיקת מספר הערים בעלות מספר הנדבקים הכי גדול מקובץ region. | מקרא: 0- עיר לא משלושת הערים הגדולות ביותר  1- עיר משלושת הערים הגדולות ביותר |

**הנחות בעת בניית המודל**

* **הנחנו שטבלת הroute מדויקת ולכן כאשר לא היה מידע אודות ביקורים של חולה הצבנו את מספר הערים אשר ביקר כ- '0'.**
* **בתת מחוז/עיר בו איננו יודעים את מספר החולים הגדרנו כ- '0'.**
* **הנחנו כי כל המטופלים ללא מועד אישור הדבקה, מידע על עיר/תת מחוז מגוריהם יושמטו עם כלל המידע לגביהם מהנתונים.**
* **כיוון שהיה ויכוח בפורום בנוגע למה נחשב כ"עיר" החלטנו להשתמש בשילוב של העיר והמחוז כיוון ששמנו לב כי העמודה שקיבלנו מה**CSV **בשם** Code **הכילה את הקוד המנהלי הייחודי של כל תת מחוז/עיר ולכן תייצג בצורה מיטבית את קושי בהגדרה של "עיר" ולא תהרוס את הנתונים. בחרנו בגישה זאת שכן לדעתנו אם היינו מאחדים ערים או עובדים לפי המחוז(province) מידע חשוב היה מושחת.**
* **השארנו את עמודה infected\_order** בטבלת העובדות שכן לדעתנו היא יכולה לשמש לשאילתות רבות אשר ידרשו בהמשך כדי להתחקות אחרי רצף ההדבקה(ידוע כקריטי מאוד בהתמודדות עם הווירוס).
* **השארנו את עמודה tests** בטבלת מימד התאריך שכן לדעתנו היא יכולה לשמש לשאילתות רבות אשר ידרשו בהמשך כדי לעקוב או כדי להציג לציבור את מספר הבדיקות שבוצעו ברחבי כלל קוריאה.
* **ערכי Outliners נוקו וטופלו לפי חשיבות העמודה וחשיבות הנתונים.**

חלק ב הסבר(קיים באנגלית גם בקובץ jupiter)

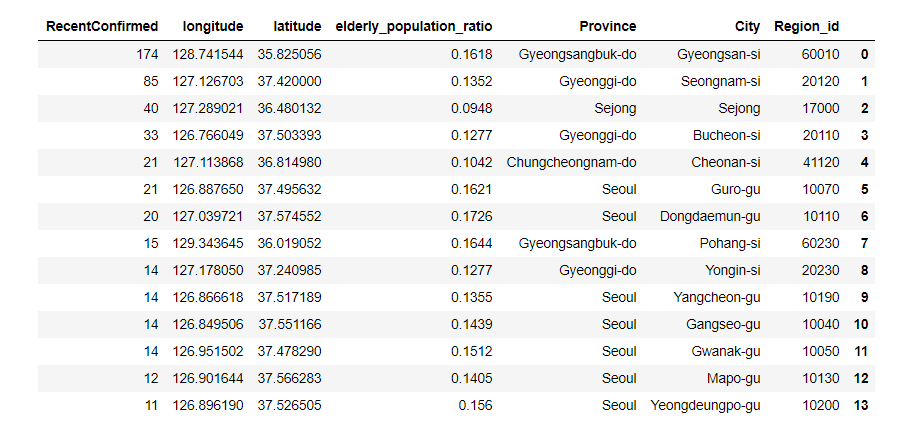
**מענה שלנו לגבי הדרישה:**

"צור מידע שיעזור לך להחליט אילו ערים בדרום קוריאה צריכות להיסגר כדי להפחית את התפשטות הנגיף."

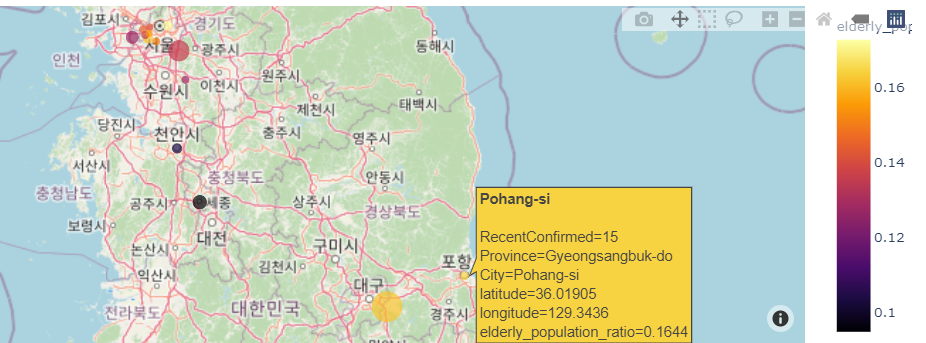
**לדעתנו על מנת להחליט אילו ערים על קוריאה להטיל סגר בדחיפות עלינו להביט במשתנים אשר ייחודים לווירוס אשר אתו אנו מתמודדים. וירוס הקורונה מתאפיין(בעת כתיבת תרגיל זה) ב:**

* **נראה כי אנשים מבוגרים, קשישים ואנשים הסובלים ממצבים רפואיים קיימים (כמו אסתמה, סוכרת, מחלות לב) פגיעים יותר לחלות קשה בנגיף.**
* **וירוס הcovid-19 מתאפיין בהדבקה מאוד חריפה והתפשטות מהירה לאזורים חדשים ובנוסף לכך הוא מתאפיין בתקופת דגירה ארוכה לכן נדרשת תגובה מהירה מאוד כאשר מזהים אזור אשר נדבק , כדי למנוע הדבקה לאזורים נוספים.**

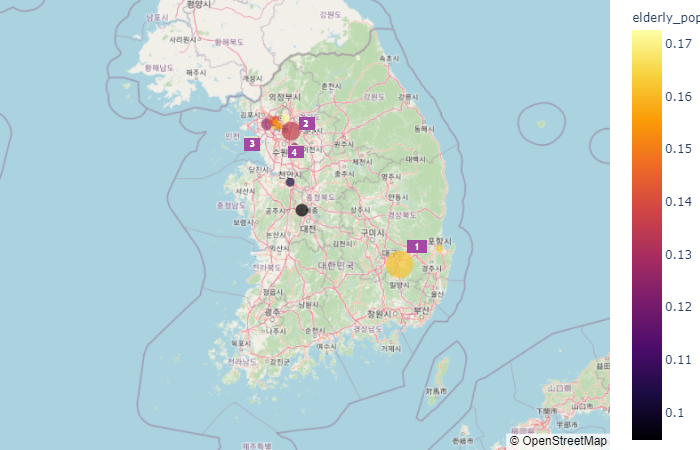
**בחרנו למקד את השאילתה שלנו בחיפוש אחר הערים אשר מכילות את אחוז הקשישים הגבוה ביותר בשילוב עם מספר ההדבקות אחרונות בשילוב נוסף עם תקופת הדגירה הממוצעת של הוירוס(שבועיים) ולכן השאילתה שלנו:  
*"סכום החולים הפעילים בכל אזור(עיר, מחוז) שאושרו מהתאריך 2020-03-05 והילך (טווח הזמן שמתחילים התסמינים) שיש בו יותר מ 20 חולים."***

**התוצאה שקיבלנו נתנה לנו מספר ערים שאכן הכילו חולים חדשים:**

**החלטנו שאין זה מספיק כדי להגיע להחלטה לגבי אילו ערים בסיכון גבוה יותר ולכן החלטנו להשתמש בPlotly על מנת להציג את הערים הנ"ל על גבי מפה אינטרקטיבית אשר תציג כל מרכז הדבקה(ואת פרטיו) בשילוב עם מספר הנדבקים(גודל העיגול) ואחוז הקשישים(הצבע).**



**גרף 1: מפת דרום קוריאה אינטראקטיבית תצוגה של שאילתה 5**

****

**גרף 2: מפת דרום קוריאה -עזר להסבר מסקנות**

**החלטנו כי התת מחזות הכי דחופים לסגירה הינם:**



* המחוזות המזרחיים **Gyeongsan-si, Pohang-she** (מסומנים ב1) הראו בבירור אחוז גבוהה של אוכולוסייה קשישה ומספר גדול יחסית של נדבקים חדשים ולכן נמליץ לסגור את המחוזות הדרום מזרחיים כליל.
* סביב העיר סיאול ניתן לראות אחוז גבוה של כיסי הדבקות ולכן נמליץ להטיל סגר כללי על סיאול על כלל תת מחוזותיה ושכונותיה.(מסומנת ב2)
* הערים אשר נמצאות בטווח קרוב לסיאול לדעתנו צריכות גם כן להסגר, בדחיפות גבוה היא עיר הנמל " **Incheon" ו**"**Suwon**" (מסומנות ב3,4)
* נדגיש כי לדעתנו את כלל הערים המופיעות בשאילתה צריך להכניס לסגר כולל ערים כגון "Sejong" אשר בעלת אחוז קשישים נמוך, אך בעדיפות נמוכה יותר.

\*נדגיש כי על מנת להפעיל את הויזואליזציה במחברת jupiter צריך להריץ את כל הקוד.  
ניתן לצפות בה ללא הרצה בקובץ HTML שמצ"ב להגשה.