נדב גרינברג 305780504
גיא זקס 203574009

מדעי הנתונים - עבודה 4

דוח סיכום מטלה

המערכת מציגה ממשק למשתמש - בו יוכל לטעון קובץ נתונים, להכין את הערכים ולבצע קיבוץ לאשכולות. המשתמש יוכל לקבוע את מספר האשכולות ומספר האיטרציות בה ירוץ המודל KMeans על קובץ הנתונים הנבחר.

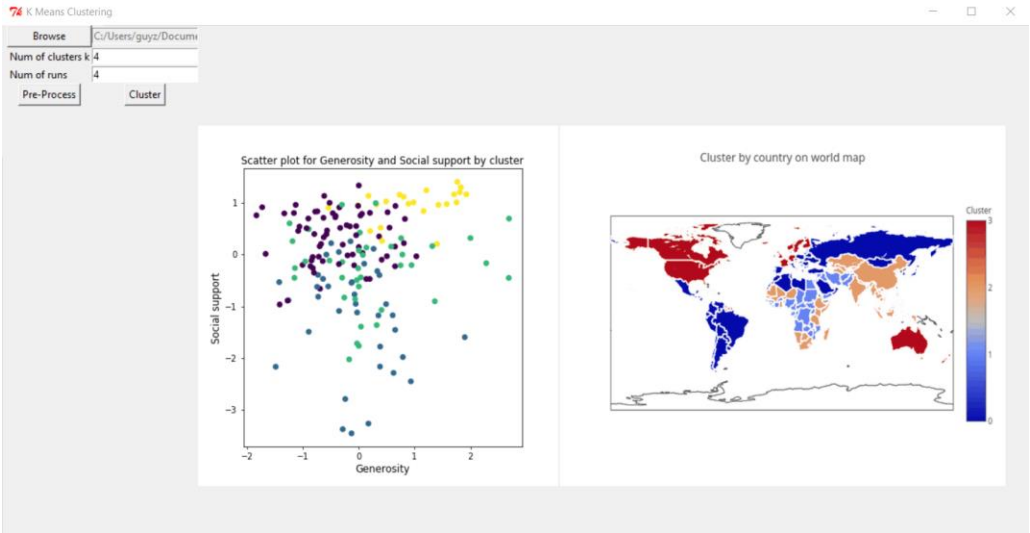
שפת מימוש: python 2.7.

חבילות בשימוש: pandas, sklearn, tkinter, sys, matplotlib, plotly

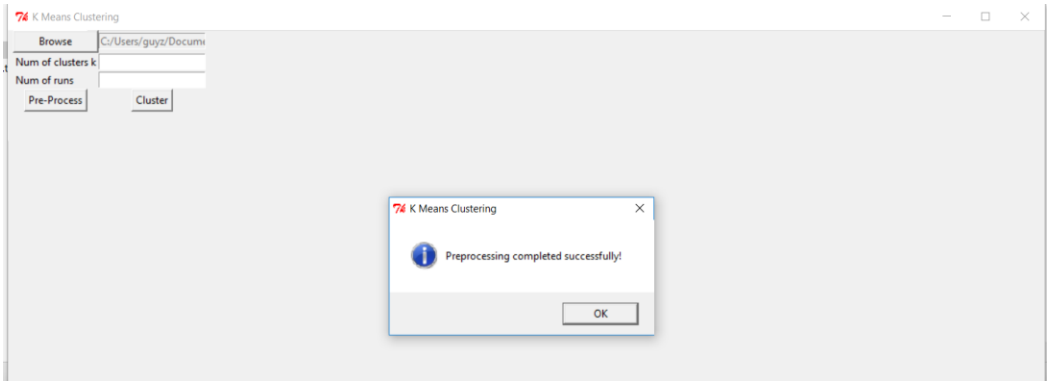
ממשק המשתמש

הממשק נכתב בעזרת החבילה Tkinter – בה השתמשנו כדי להגדיר את מיקומי האלמנטים במסך, קישור בין כפתורים ותיבות טקסט לפונקציות בקוד ומעבר בין מצבים שונים במערכת.

הממשק לאחר סיום הקיבוץ:



הממשק לאחר ביצוע ה-pre-processing:



נדב גרינברג 305780504
גיא זקס 203574009

מדעי הנתונים - עבודה 4

<u>ארכיטקטורה</u>

הקוד בנוי מ-2 קבצי py:

- main.py - מנהל את התהליך הרציף בקוד, מכיל את פונקציות ה-GUI, תקינות קלט וביצוע ה-clustering
- cleanData.py – מכיל את כל פעולות ניקוי הנתונים הדרושות כדי להריץ את המודל

פונקציות ב-main.py:

- browseFile – בפונקציה זו ממומש תהליך פתיחת חלון לבחירת קובץ הנתונים לטעינה.
- preprocessing – פונקציה זו שולחת את הקובץ הנבחר ל-cleanData.py, ובעת חזרת הנתונים הנקיים מציגה בממשק חלון להצלחת סיום התהליך או חלון שגיאה במקרה שהתהליך כשל.
- clustering –
  - הפונקציה מאמתת את תקינות הערכים שהתקבלו מהמשתמש (מספר שלם חיובי). במקרה של שגיאה – יוקפץ חלון למסך עם השגיאה הרלוונטית שנמצאה
  - מימוש אלגוריתם KMeans באמצעות החבילה הרלוונטית ב-sklearn. האלגוריתם מחזיר את ערך האשכול של כל רשומה (כאשר כל רשומה נפרדת מייצגת מדינה אחת, כפי שיפורט בהמשך). ערך האשכול מתווסף ל-dataframe של כלל הנתונים על המדינות.
  - תצוגת התוצאות על 2 גרפים:
    - Scatter – לפי ערכי Generosity ו-social support. ממומש ע"י החבילה matplotlib, כאשר ערך x הוא הערכים המנורמלים של Generosity וערך y הוא הערכים המנורמלים של social support. צבע הנקודות בגרף הוא לפי ה-cluster שהאלגוריתם חישב.
    - Cluster on map – ערכי הcluster בתצוגת של מפת העולם. ממומש ע"י החבילה Plotly. כאשר המדינות ממוקמות על המפה והצבעים נקבעים לפי הערך שהחזיר האלגוריתם. לצורך הצגת גרף זה בממשק, התוצאה נשמרה כקובץ תמונה במחשב והועלתה לממשק
- תצוגת הממשק והאלמנטים שבו

פונקציות ב- cleanData.py:

- הקובץ מכיל פונקציה אחת שבה מבוצעים:
  - קריאת קובץ ה-excel שהמשתמש הזין ע"י חבילת pandas
  - עבור הנתונים המספריים בקובץ הנתונים:
    - מילוי חוסרים לפי הממוצע של העמודה – ע"י פונקציות apply, fillna של pandas
    - המרת הנתונים לערך הנורמלי-סטנדרטי – ע"י הפונקציה standardScaler של החבילה sklearn
  - הנתונים המספריים החדשים מתחברים לנתונים הטקסטואליים בקובץ. על קובץ הנתונים החדש נעשית פעולת groupby שממומשת ע"י החבילה pandas, ובה קיבוץ הממוצע של כל עמודה לפי המדינה.
  - החזרת ערך בוליאני אם הפעולה הסתיימה בהצלחה וקובץ נתונים המכיל רשומה אחת עבור כל מדינה – עליו נבצע את פעולת ה-clustering