Analytics of big data - Targil 1
הגשה עד 19/5/2020

המטרה של התרגיל הזה הוא ללמוד עבודה עם דאטה וספארק.
התרגיל יתבסס על נתונים שקשורים להתפרצות הקורונה, התפרצות שפעת עונתית בקשר למניות (אפשר להסתכל על אינדקס של הבורסה או על מניות ספציפיות למשל גוגל)
להלן אתרים אפשריים, מומלץ לבחור אחרים.
https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide
https://healthdata.gov/dataset/deaths-pneumonia-and-influenza-pi-and-all-deaths-state-and-regi(on-national-center-health, https://finance.yahoo.com/quote/%5EIXIC/history?p=%5EIXIC
https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset
יש מקום ליצירתיות בתרגיל, בבחירות הדאטה, בשאלת המחקר, מציאת קורלציות מעניינות, להצגת התוצאות ועוד.
המטרה היא להראות קורלציות מעניינות בין המאגרים השונים

ויזואליזציה של הדאטה

- הציגו התפלגויות של כל דאטה ביחס לשתי עמודות מעניינות (למשל מוות מקורונה).
- הציגו גרף המציג את הגידול/שינוי של כל דאטה לאורך זמן והשוו ביניהם.
- הציגו יחסים מעניינים בין עמודות (או סכום של עמודות לאורך זמן).
- בחרו 2 עמודות והציגו עבורם box plot, הסבירו את התוצאות.
- בחרו שתי עמודות והציגו עבורם density plot, הסבירו את התוצאות
- בחרו שתי עמודות והציגו עבורם scatter plot, הסבירו את התוצאות
- בצעו אגרגציה על פי מיקום (כלומר מדינה) והציגו את המידע על ידי google maps api

בחירה של מאפיינים והכנה של דאטה מסכם

1. צרו טבלה המאחדת את הטבלאות שבחרתם שבה יש שורה אחת עם כל הפרטים (העמודות) המעניינות לבניית מודל.
2. הציגו גרפים המשווים בין עמודות שונות (נסו לחפש קורלציות)

בנית מודל

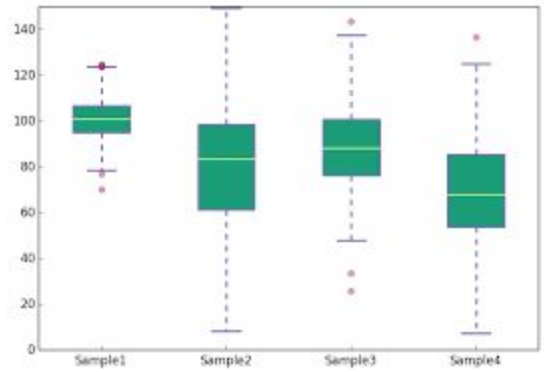
1. מהם המשתנים שאתם חושדים שיש ביניהם קורלציה?
2. השתמשו במודל של רגרסיה לינארית והריצו על המשתנים הנ״ל, הסבירו את התוצאות.

1.

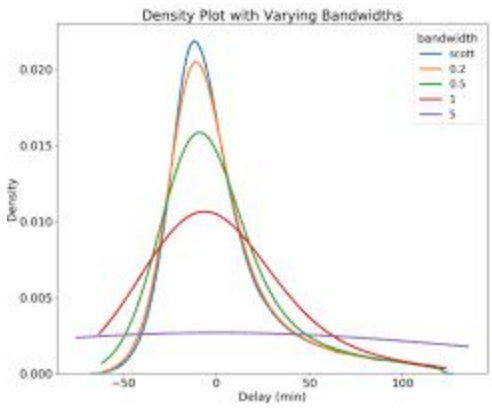
הוראות הגשה:

1. יש לשלוח לבודק את קבצי ה-JUPYTER עם ההסברים והתוצאות בתוך הקובץ.
2. יש להעביר הרצאה של 5-10 דקות בכיתה בשיעור של ה

Box plot



Density plot



Scatter plots