TEL AVIV UNIVERSITY אוניברסיטת תל-אביב

משימה 1

הפקולטה להנדסה
המחלקה להנדסת תעשייה וניהול
מבוא למדעי הנתונים (Introduction to Data Science)
-סמסטר א' תשע"ז

כללי:

- מטרת הפרויקט היא לאפשר תרגול מעשי בפועל של עקרונות שיטת CRISP-DM לניהול מחזור החיים הכולל של פרויקט אנליטיקה עסקית בסביבת R .
הפרויקט כולל תרגול של מספר מודלים של כריית נתונים כפי שנלמדו בכיתה. התרגול נועד לאפשר להיחשף לכל שלב בשיטה ומחייב תיעוד מפורט של הפעולות אשר נעשו בכל שלב במחזור החיים של הפרויקט.

- במשימה זו, יעשה שימוש בנתוני דירוגי הסרטים של IMDB הנקרא:
"Movie information and user ratings from IMDB.coms".
מטרת המשימה היא להציג את היסודות של שפת התכנות R תוך תרגול של שני מרכיבים מרכזיים של המודל CRISP-DM: pre-processing ו- data exploration.

קווים מנחים:

נקודות תורדנה על חוסר הצמדות לקווים המנחים:

- השתמש בקובץ קוד המופיע במודל. **אין להסיר ממנו תגובות!!!**
- שאלות ייענו בפורום **בלבד.**
- בבדיקה יינתן דגש על יעילות הקוד אז שימו לב.
- הקוד להיות מלווה בהסברים לצד כל שורה.
- הניקוד יופיע בתחילתה של כל שאלה בסוגריים והוא יוכל לספק אינדיקציה לרמת הקושי שלה (כך שאם אתם נתקעים, נסו להמשיך הלאה).
- הציון הסופי שלכם יושפע גם מ'קריאות' הקוד שלכם, אז שימו לב לבצע זאת באופן קצר ומתומצת והשתמשו בשמות משמעותיים עבור המשתנים.

הגשה:

- **מועד אחרון להגשה:** 19/12/2016
- **קבוצות:** קבוצות של 2 סטודנטים
- **אופן ההגשה:** קובץ R שיועלה לאתר המודל (moodle) על ידי **אחד** מחברי הקבוצה. לא לשכוח להכניס ת.ז של כל חברי הקבוצה בראש הקוד
- **ניקוד:** 1-100 , קוד שלא ירוץ יזכה ב-0 נקודות
- **מייל עוזר הוראה:** omermiran@gmail.com

טיפים מומלצים:

- במהלך התרגול נדרש להתקין חבילות רבות לאורך כתיבת הקוד. בכדי להתקין חבילה ב R יש להשתמש בפקודה `install.packages` (חפשו את הפקודה בגוגל)
- קבלת עזרה על פונקציה מסויימת- תפעיל את הפונקציה "?" לפניה, לדוגמא:
`?nameOfFunction` , בתחתית עמוד העזרה (help page) שייפתח תוכלו למצוא בדרך כלל דוגמאות טובות לשימוש הפונקציה. תוכלו להתשמש בדוגמאות אלו ולשנות בהתאם לצורכיכם.
במידה ואין לכם את החבילה של אותה פונקצייה, לא תראו עמוד עזרה לפונקצייה ותצטרכו להתקין את החבילה כמו שהוסבר בסעיף הקודם.

ה -Data set מצורף במודל.
תוכלו למצוא הסבר לקובץ הנתונים דרך הלינק הבא:
https://vincentarelbundock.github.io/Rdatasets/doc/ggplot2/movies.html

בהצלחה!!!

הפקולטה להנדסה
המחלקה להנדסת תעשייה וניהול
מבוא למדעי הנתונים (Introduction to Data Science)
-סמסטר א' תשע"ז

1. Loading the data:

(0) **1.a.** הורידו וחלצו את קובץ המידע (movies.csv) מהמודל לתקייה מקומית ייעודית לעבודה זו.

(2) **1.b.** קבעו את ספריית העבודה שלך להיות התקייה מהסעיף הקודם למען גישה נוחה. מעתה ואליך, אין צורך להשתמש בנתיב המלא אלה רק את שם הקובץ בתקייה זו

(3) **1.c.** ייבאו את הקובץ movies.csv ל-R ושמור אותו בשם "data". שימו לב שלנתונים בקובץ כבר יש מספרי שורה (בידקו את הארגומנטים בפונקצייה על מנת לבצע את הייבוא באופן טוב)

(1) **1.d.** העמודות "r1","r2",…,"r10" לא יהיו רלוונטיות עבורינו, לכן הסירו אותם מה data.

(3) **1.e.** הצג את שמות כל העמודות ואת סוג הנתונים שכל אחת מהן מכילה בעזרת פונקציה אחת.

2. Data exploration:

(2) **2.a.** השתמשו בפקודה שמציגה מספר מוגבל של השורות הראשונות. בדקו שהפונקציה שהשתמשתם מראה את הנתונים בצורה טובה ובכך תבין האם טענת את הקובץ בצורה טובה.

(3) **2.b.** הנתונים מכילים סרטים רבים עם מעט מאוד הצבעות. מחק את כל הסרטים שלהם יש פחות ממאה הצבעות (votes)

(4) **2.c.** משתנים קטגוריים: אפשר לראות (לפי סעיף 1.e) כי הטורים: "short"…,"Animation","Action" כולם מסוג Integer, אך הם צריכים להיות משתנים לוגיים (True/False). הפוך את הפיצ'רים הללו ללוגיים. מדוע אי אפשר לאחד את כל הפיצ'רים האלו לפיצ'ר אחד שנקרא "Genre"

(5) **2.d.** משתנים נומריים: צרו תקציר סטטיסטי (summery statistics) ובו: ממוצע, סטיית תקן, שונות, מינימום, מקסימום, חציון, טווח וחמישון (mean, standard deviation, variance, min, max, median, range, and quantile). בצע זאת לכל המשתנים הנומריים

3. Missing values:

(3) **3.a.** ספרו את כמות השורות בהן יש ערכים חסרים, ציין את שם הפיצ'ר בעל כמות הערכים החסרים הגדולה ביותר.

(6) **3.b.** על פי האסטרטגיות שנלמדו, האם זה הגיוני להוריד את כל השורות אשר יש בהן ערך חסר?
אם כן, הסירו את כל השורות עם ערך חסר
אם לא, מצאו דרך אחרת להתמודד עם הערכים החסרים ויישמו אותה (לא חייב להיות מסובך).

הפקולטה להנדסה
המחלקה להנדסת תעשייה וניהול
מבוא למדעי הנתונים (Introduction to Data Science)
-סמסטר א' תשע"ז

4. Data normalization:

(4) **4.a.** צור qq-plot לכל אחד מהפיצ'רים: 'year', 'rating', 'votes'. הסבר מה אפשר ללמוד מ qq-plot לגבי ההתפלגות של פיצ'ר

(7) **4.b.** על פי הqq-plots, האם יש צורך לנרמל את שלושת הפיצ'רים? אם לא, איזה פיצ'רים צריך לנרמל? האם זה הגיוני שנתונים אלה לא מתפלגים נורמלית? באיזה שיטה הייתם מנרמלים כל פיצ'ר? צרו טור חדש למשתנה data לכל פיצ'ר שנירמלתם עם הנתונים המנורמלים.
לדוגמא: אם נרמלתם את הפיצ'ר votes צרו טור חדש עם השם votes.norm שבו הנתונים המנורמלים.

(3) **4.c.** ציינו 3 דרכים לנרמול נתונים והסבירו מדוע חשוב לבצע זאת

5. Outlier detection:

(2) **5.a.** צרו גרף אחד המכיל את כל הbox plots לכל אחד מהמשתנים הנומריים (למעט הפיצ'ר/ים שבחרת לנרמל, במקומם הצג את ה box plot של המשתנים המנורמלים שלהם)

(4) **5.b.** צרו box plot לכל פיצ'ר בנפרד ופרט על 2 עובדות מעניינות שנובעות מהגרפים. (מגמות וכדו')

(7) **5.c.** אפשר לראות כי בפיצ'ר 'length' יש הרבה נק' החשודות כ outliers. הוציאו את נקודות אלו מהdata

(7) **5.d**. השתמשו במדד ה LOF בכדי להוציא outliers כאשר אתם משתמשים בפיצ'רים:
"rating","length","votes". השתמשו בk=20 והוציאו את כל הנקודות אשר ערך הLOF שלהם מעל 1.5

6. More exploration and visualization:

(5) **6.a.** הציגו גרף עמודות אשר מראה את כמות הסרטים לכל סוג (genre) הציגו כל עמודה של סוג סרט בצבע שונה

(12) **6.b.** אורך הסרט וכמות הדירוגים שהוא קיבל משפיעים מאוד על הציון שהוא מקבל.
צרו וקטור עם 3 רמות של אורך סרט (length): Long>110/ 90<=Medium<=110 / Short< 90
צרו וקטור עם 2 רמות של כמות דירוגים (votes): (Many votes>500 / Few Votes<500)
חשבו את הדירוג הממוצע של כל אחד מהתת קטגוריות שיוצרות הרמות.
לדוגמא, ממוצע הציונים של סרטים ארוכים עם מעט דירוגים. (ישנן 6 פרמוטציות)

(5) **6.c.** הציגו 2 density plots:
1. density plot של דירוגים כתלות באורך הסרט (לפי 3 הרמות שיצרתם בסעיף הקודם)
2. density plot של דירוגים כתלות בכמות הדירוגים לאותו סרט (לפי 2 הרמות שיצרתם בסעיף הקודם)

7. Correlation analysis:

(7) **7.a.** הציגו correlation plot של כל הפיצ'רים הנומרים. שימו לב שהגרף ברור.

(5) **7.b.** מצאו את הפיצ'רים שקיים בינהם מתאם של מעל 0.5 ופיצ'רים שיש בינהם מתאם של מתחת ל0. האם אפשר לראות מגמה ברורה? האם אפשר "להיפטר" מאחד הפיצ'רים?