**עריכת/הכנת הנתונים - Data Preparation**

יבואויצואהנתונים –Data import & export

פורמטים של הקבצים-  
לקרוא את הנתונים מ ולכתוב אותם אל-  
פורמטים של R  
קבציCSV  
קבציEXCEL  
מסדינתוניםODBC  
מסדינתוניםSAS

save.image () - לשמור את סביבת העבודה הנוכחית לתוך קובץ, שומר כל דבר.  
ReadRDS ()– לקרוא אובייקט R בודד מתוך קובץ  
saveRDS ()– לשמור אובייקט R בודד לתוך קובץ  
יתרונות של readRDS () ו- saveRDS (): אתה יכול לשחזר את הנתונים תחת שם אובייקט שונה.  
יתרונות של load() ו- save(): ניתן לשמור אובייקטים של R רבים בקובץ אחד.

**עריכת הנתונים**

נתונים ב0עולם האמיתי הם מלוכלכים.  
שלמות: חסרי ערך של התכונות, חוסר מאפיינים מסוימים מעניינים, או מכילים רק מידע כוללני.  
רועש: מכיל שגיאות או חריגים  
עקביות: מכילסתירותבקודיםאובשמות  
  
אין נתונים איכותיים, אין תוצאות איכותיות של הכרייה!  
החלטות איכותיות חייבות להיות מבוססות על נתונים איכותיים.   
מחסני נתונים צריכים לשלב באופן עקבי נתונים איכותיים.  
הערכה של נתונים איכותיים משקף על אמון בתוצאות.  
  
הכנת הנתונים לניתוח  
תחשוב על הנתונים שלך:  
– איך זה נמדד, מה המשמעות של זה?  
- ערכי או קטגורי  
- סידורי : לדרגה יש משמעות (לערך מספרי לא בהכרח. רמתהשכלה, דרגהצבאית)  
- ערך שלם (למרחקים בין מספרים שלמים יש משמעות. טמפרטורה, זמן)  
- יחס (לאפס יש משמעות. שברים ויחסים הגיוניים. כסף, גיל, גובה)  
זה אולי נראה מובן מאליו מהו ערך הנתונים הנתון אבל לא תמיד (מדדכאב, דירוגיסרט, וכו')

משימותבעיבודמקדיםשל הנתונים - Tasks in Data Preprocessing

* ניקוי נתונים: בדיקה עבור נתונים איכותיים, נתונים חסרים.
* שינוי נתונים: נרמול וצבירה
* צמצום נתונים: מקבל ייצוג מופחת בהיקפו אבל מייצר את תוצאות אנליסטיות דומות או זהות
* Data discretization–שילוב של צמצום ושינוי אבל עם חשיבות מיוחדת, במיוחד עבור נתונים מספריים.

ניקוי הנתונים / איכות הנתונים - Data Cleaning / Quality

* מדידות פרטיות – רעשים רנדומליים במדידות פרטיות (חריגים, נתונים רנדומליים מכניס שגיאות, רעש במשימות מותג, יש אפשרות לתקן או להחליק החוצה)
* טעויות שיטתיות
* מידע חסר – חסר באופן רנדומלי (שאלות בשאלון שהנשאלים שכחו באופן רנדומלי למלא אותם), חסר באופן שיטתי (שאלות שאנשים לא רוצים לענות עליהן, מטופלים שחולים מידי כדי למלא מבחן)

מידע חסר

* נתונים לא תמיד זמינים – הרבה רשומות אין ערכים רשומים עבור כמה תכונות (משתתפים בסקר, מקורות שונים של נתונים)
* נתונים חסרים יכולים לקרות בגלל – תקלה בציוד, המידע לא הוכנס כראוי, מידע לא זמין, גרסאות שונות של מידע אשר מוזגו יחד.

איך להתמודד עם נתונים חסרים?

* להתעלם מהמשתנים – אפשרי רק עבור אחוזים קטנים של ערכים חסרים.
* להשתמש בקבוע גלובלי (כמו ממוצע משתנה) כדי למלא את הערך החסר – "לא ידוע" כקטגורי, לנתונים רציפים זה יקטין משמעותית את השונות.
* להשתמש בערך אקראי כדי למלא את הערך החסר– שומר על השונות, לא מזיק.
* להשתמש בזקיפה – השכן הקרוב ביותר, מבוסס מודל (מבוסס רגרסיה או בייסן)  
    
  מה אני בוחר למצב הנתון?

מה שאתה עושה תלוי ב –

* הנתונים – עד כמה הוא חסר? האם אלו ערכים "חשובים"?
* המודל – האם יכול להתמודד עם ערכים חסרים?
* האם הנתונים חסרים אקראית?
* אין תשובה נכונה
* תמיד תבדוק את החוסן של התוצאות

נתונים רועשים

* רעש – שגיאות אקראיות או שונות במשתנים נמדדים.
* ערכי תכונות שגויים (חריגים) נגרמים בשל: אוסף נתונים שגוי, בעיות בהזנת הנתונים, הגבלות בטכנולוגיה.
* בעיות נתונים אחרות הדורשים ניקוי נתונים: רשומות כפולות, נתונים חלקיים, נתונים לא עקביים.

שינוי נתונים -Data Transformation

* יכול לעזור בהפחתת ההשפעה של ערכים קיצוניים
* הפחתת השונות – לעיתים קרובות מאוד שימושי כאשר מתמודדים עם נתונים מוטים (הכנסות). שורש ריבועי, דו צדדי, לוגריתמים, להעלות בחזקה. Logit - המרות הסתברויות בין 0 ל- 1 לקו-אמת
* נורמליזציה - מדורגליפולבטווחקטן, שצוין. לעיתים אנו רוצים לקבל את כל המשתנים באותו קנה מידה. נרמול מינימום-מקסימום
* תכונה/ בניית תכונה – תכונות חדשות שנבנו מאלה שניתנו.

התמודדות עם נתונים מאסיביים - Dealing with massive data

מה קורה אם הנתונים לא מתאימים למחשב שלי (או קריסות R)?  
מדגם – יש להקפיד לעשות אקראיות נכונה, ריבוד  
מצא שאלה קטנה – השתמש בכלים כדי להפחית את בסיס הנתונים ולנסח מחדש את השאלה.  
השתמש במסדי נתונים –MySQL היא טובה וחינמית  
לחקור אסטרטגיות לצמצום נתונים – יכול להפחית n או p

הפחתת נתונים – ממדי ההפחתה - Data Reduction: Dimension Reduction

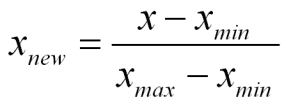
* באופן כללי, כרוך באובדן של מידע אודות x.
* אם P הממדית היא מאוד גדולה (למשל, 1000), המייצגת את הנתונים בשטח נמוך-ממדי עשוי להפוך את הלמידה לאמינה יותר, למשל, דוגמת האשכולות: 100 נתונים ממדיים, אםמבנההאשכולקייםרקב 2 מהממדים, הממדים האחריםהםרקרעש. אם שאר ה-98 הם רק רעש (יחסית למבנה אשכול), אז האשכול יהיה קל יותר לגילוי אם רק מתמקדים במרחב d2.
* הפחתת ממדים יכולה גם לספקפרשנות / תובנה
* בחירת תכונות- השתמש ב-EDAכדי למצואמשתניםחסריתועלת. השתמש בחיפוש ממצה על מודל פשוט (למשל רגרסיה). השתמש בשיטות היוריסטיות כמו שיטות בשלבים.

הפחתת נתונים: דגימה

* אל תשכחו את הדגימה
* בחר נציג של התת קבוצה של הנתונים (דגימה אקראית פשוטה יכולה להיות בסדר אבל היזהרו ממשתנים מוטים)
* שיטות דגימה מרובדת: אומדן האחוז של כל מחלקה במסד נתונים הכללי, בשילוב עם נתונים מוטים, נטיית התוצאות יכולה להיות שימושית אם התגובה לא מאוזנת.

שינוי קנה המידה של הנתונים - rescaling data

Normalization– נרמול – משנה את קנה המידה של המשתנים המספריים בטווח [1, 0].



Standardization– תקינה, סְטַנְדַּרְטִיזַצְיָה- לשנות את זה לאפס ממוצע ושונות יחידה.

