# The election challenge

1. בשלב ראשון טענו את המידע והורדנו את כל השורות בהן היו חסרים ערכים.
הרצת describe הראתה לנו שיש שדות עם ערכים שליליים, שלא הגיוני שיהיו שם. לאחר שהורדנו גם את כל השורות עם הערכים השליליים נשארנו עם 8146 שורות.
2. Feature selection- מאחר והמטרה היא לנקות פי'צרים בצורה שמרנית, החלטנו לחפש לכל פיצ'ר סיבות שונות להשאיר אותו, ואת הפיצ'רים שלא מצאנו סיבה להשאיר, נוריד.
התחלנו עם מבחן chi2 לערכים הנומינליים, ומבחן ANOVA לערכים המספריים.
החלטנו לשמור את הפיצר'ים שה-p-value שלהם קטן מ0.05.
לאחר מכן הדפסנו את פונקציות הצפיפות של המשתנים הרציפים, וההיסטוגרמות של המשתנים הדיסקרטיים, ולמדנו ש:
   - בפיצ'ר Last_school_grades כל הערכים מ-60 ומעלה מתנהגים אותו הדבר, לכן אפשר לצמצם אותם לערך אחד.
   - הערכים Avg_monthly_expense_when_under_age_21
   Avg_Residancy_Altitude תלוים לינארית אחד בשני, לכן אפשר להוריד אחד מהם, אבל לפני זה כדאי להשתמש בו להשלמת ערכים עבור השני. במימוש הסופי הורדנו עוד שדות שתלויים לינארית.
   - יש התאמה בין ההיסטוגרמות של Last_school_grades ו Most_Important_Issue.
   - יש פיצ'רים שמתפלגים פואסונית (כמו Num_of_kids_born_last_10_years) ויש שמתפלגים (בקירוב כמובן) אקספוננציאלית (כמו AVG_lottary_expanses) אבל החלטנו לא להתייחס אליהם בצורה מיוחדת.
   - מסקנות נוספות נמצאות בnotebook לצד הגרפים הרלוונטים.
3. האלגוריתם:
   a. מנקים את המידע.
   b. מחפשים פיצ'רים דיסקרטיים עם התאמה בין הערכים, משלימים ערכים חסרים וזורקים אחד מהם (אם רק אחד מהם נומרי נעדיף לשמור אותו).
   c. מחפשים פיצ'רים נומריים עם קורלציה לינארית גבוהה בכל הלייבלים, משלימים ערכים חסרים וזורקים אחד מהם.
   d. בוחנים את הפיצ'רים הנומינלים בעזרת chi2 ואת המספריים בעזרת ANOVA (פיצ'רים שמתפלגים לינארית הופכים קודם כל לנורמלים). ניסינו גם להריץ ANOVA על הlog values אבל לא קיבלנו פיצ'רים נוספים שכדאי לשמור.
   e. מריצים SFS עם מספר מסווגים שונים, אם יש פיצ'ר שנבחר ע"י אחד מהם נוסיף אותו לבחירה שלנו. (היחיד שנבחר הוא gender, אנחנו לא בטוחים לגביו אבל מכיוון שעדיף להשאיר פיצ'רים מיותרים מאשר לזרוק פיצ'רים חשובים אז נשאיר אותו)

בפועל הרצנו כמה פעמים ובכל פעם הפיצ'רים שהתווספו בגלל SFS היו שונים (בין היתר קיבלנו את Avg_education_importance, Occupation, Occupation_Satisfaction, %Time_invested_in_work), ומהסתכלות בגרפים שלהם נראה שההבדלים בין הלייבלים די מקריים, לכן אנחנו חושבים שעדיף בלעדיהם.

מכיוון ש Number_of_valued_Kneset_membersלא חזר מאף ריצה של SFS, ומראש הוא נכנס עם p-value די גבוה ב-0.0219284964384) ANOVA) אנחנו חושבים שעדיף לוותר עליו. בסוף האלגוריתם הדפסנו השוואה של התוצאות עם כל הפיצ'רים שבחרנו לעומת התוצאות כאשר מורידים כל אחד מהפיצ'רים בקבוצת הפיצ'רים שלא נבחרו באף הרצה של SFS. גם שם אנחנו רואים שהורדה שלו לא פוגעת בתוצאות. הפיצ'רים איתם נשארנו:

'Political_interest_Total_Score', 'Yearly_ExpensesK', 'Avg_monthly_household_cost', 'Number_of_valued_Kneset_members', 'Yearly_IncomeK', 'Married', 'Last_school_grades',

'Avg_monthly_expense_when_under_age_21', 'Looking_at_poles_results', 'Financial_agenda_matters', 'Overall_happiness_score', 'AVG_lottary_expanses'

המשמעות של הפיצ'רים שבחרנו ביחס לליבל (כפי שניתן לראות מהגרפים בנוטבוק):

- Political_interest_Total_Score, Avg_monthly_household_cost, Yearly_IncomeK: בכל המפלגות התוחלת נראית דומה, אבל במפלגות מסוימות ההתפלגות מאוד צפופה סביב התוחלת (הכי קיצוני במפלגה 3, ) ובמפלגות אחרות, כמו 1,5,8 , השונות הרבה יותר גדולה. 3 הפיצ'רים האלה מאוד דומים וכנראה שאפשר לוותר על חלקם, גם Yearly_ExpensesK ו-Overall_happiness_score נראים די דומה, אבל פחות.
יתכן ש Yearly_ExpensesK, Overall_happiness_score, ואולי פיצ'רים נוספים, נגזרים מ Yearly_IncomeK עם רעש מסוים. עם המסווגים שהרצנו בסוף, הורדה של Avg_monthly_household_cost לא פגעה בציונים, לכן אנחנו חושבים שאפשר להוריד אותו. גם כשמורידים את Political_interest_Total_Score הפגיעה היא מינורית מאוד, אבל נהיה זהירים ונשאיר אותו.
- Married – רווקים נוטים מאוד להצביע למפלגות 2 ו-5, הרבה יותר מנשואים. נשואים לעומת זאת מצביעים ל4 ו-9 הרבה יותר מרווקים.
- Last_school_grades- אנשים עם ציון 30 מצביעים רק ל5,8. אנשים עם ציון 40 מצביעים רק ל 2,8. אנשים עם ציון 50 מצביעים רק ל2,5. ההצבעה בשאר הציונים מתחלקת די אחיד בין שאר המפלגות (ביחס לגודלן כמובן, הגרפים מנורמלים). כמו כן מצאנו התאמה מלאה בין כל ציון לאחת הקטגוריות ב- Most_Important_Issue. העדפנו את הפיצ'ר המספרי, אבל אולי למספרים אין באמת משמעות.
- Looking_at_poles_results- הפוך בדיוק מ- Will_vote_only_large_party. האלגוריתם שלנו הוריד במקרה את Will_vote_only_large_party אבל זו בדיוק המשמעות, מצביע מסתכל בסקרים אמ"מ הוא לא מצביע למפלגה גדולה, אמ"מ הוא יצביע בפועל למפלגה קטנה (1 או 6). בתוך כל קטגוריה ההצבעות מתחלקות אחיד בין כל המפלגות (שוב, ביחס לגודלן).
- Financial_agenda_matters – מפריד בדיוק בין מפלגות 2,5,6,8 לשאר.
- Avg_monthly_expense_when_under_age_21 – יש התאמה לינארית מלאה בינו לבין Avg_Residancy_Altitude, לכן האלגוריתם זרק את Avg_Residancy_Altitude. מהגרף רואים שהפיצ'ר מפריד טוב ל-3 קבוצות: 1, 6, וכל השאר.
מיפוי המספרים לשמות:

0=Blues
1=Browns
2=Greens
3=Greys
4=Oranges
5=Pinks
6=Purples
7=Reds
8=Whites
9=Yellows

רשימת הפיצ'רים הסופית שבחרנו:
'Political_interest_Total_Score', 'Yearly_ExpensesK', 'Married', 'Yearly_IncomeK', 'Last_school_grades', 'Looking_at_poles_results', 'Avg_monthly_expense_when_under_age_21', 'Financial_agenda_matters', 'Overall_happiness_score', 'AVG_lottary_expanses'

אנחנו מאמינים שגם ברשימה הזאת יש יתירות, אבל לשם הזהירות נשאיר אותה כך.

בנוסף, מצורף קובץ "looking at the data" ובו הגרפים שהשתמשנו בהם (קיימת גם גרסאת html לקריאה נוחה ללא צורך בהרצה).