Practical part answers:

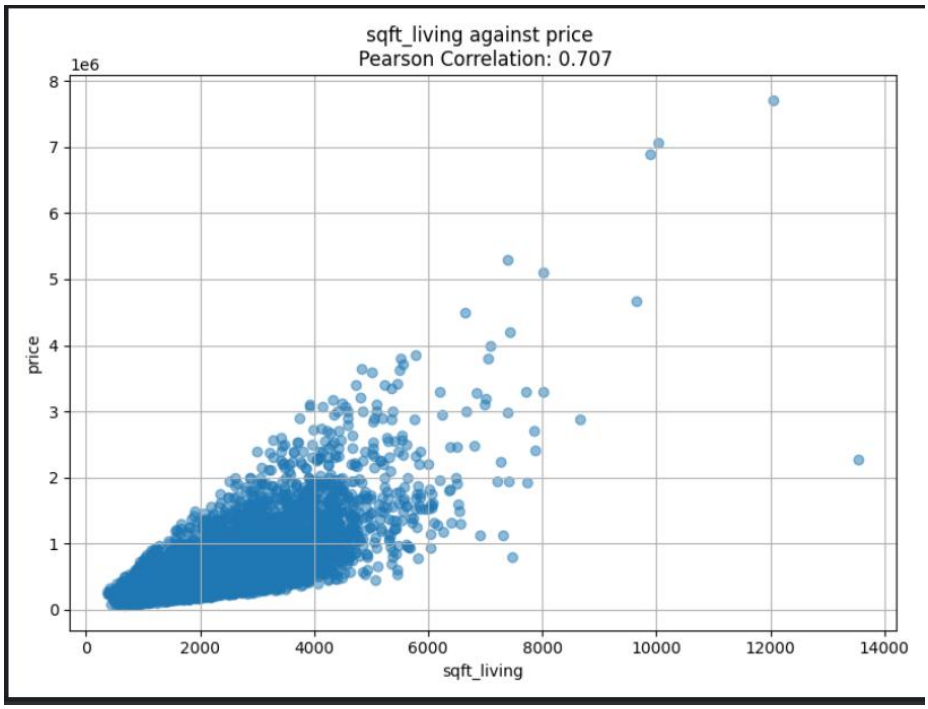
Linear Regression:

3)א) הפיצ'רים שבחרתי להוריד הם ID, DATE, YR_BUILT, YR_RENOVATED מאחר ID הוא שדה שרירותי ללא כל משמעות מתמטית על מחיר הבית וכנ"ל לגבי .DATE לגבי YR_RENOVATED, YR_BUILT הורדתי אותם כי הוספתי פיצ'רים במקומם שמחשבים את ההפרש מהשנה הנוכחית לשנה הנתונה מאחר ולדעתי אלו נתונים שיותר נוח להסתכל עליהם ולהסיק מסקנות(למשל בתים באזור גיל 30 שווים פחות שזה נתון שלא קופץ לעין בכזאת קלות לפי שנת בנייה) מאשר השנה הספציפית שבה הבית נבנה \ שופץ.

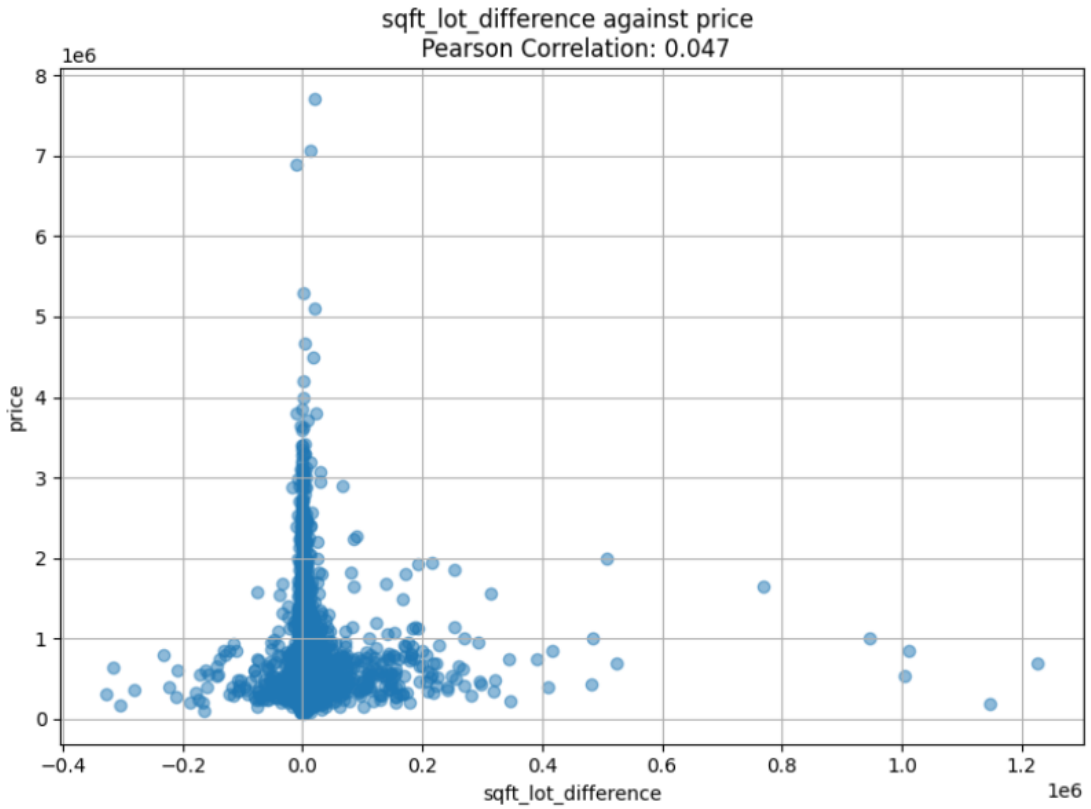
ב) הוספתי חמישה פיצ'רים חדשים: **גיל בית**, **שנים מאז שיפוץ** – נתונים יותר נוחים לניתוח לדעתי מאשר הפיצ'רים המקוריים המקבילים להם. בנוסף הוספתי **הפרש בין גודל מגורים לממוצע 15 קרובים, הפרש בין גודל מגרש לממוצע 15 קרובים.** אלה נתונים שלדעתי יתנו זווית יותר עניינית על הנתונים- איך ההפרש מהגודל הממוצע של 15 השכנים הקרובים משפיע על המחיר.

ג) לגבי נתונים INVALID/MISSING פשוט מחקתי את כל השורה מהנתונים כי לא ניתן לסמוך עליה לדעתי.

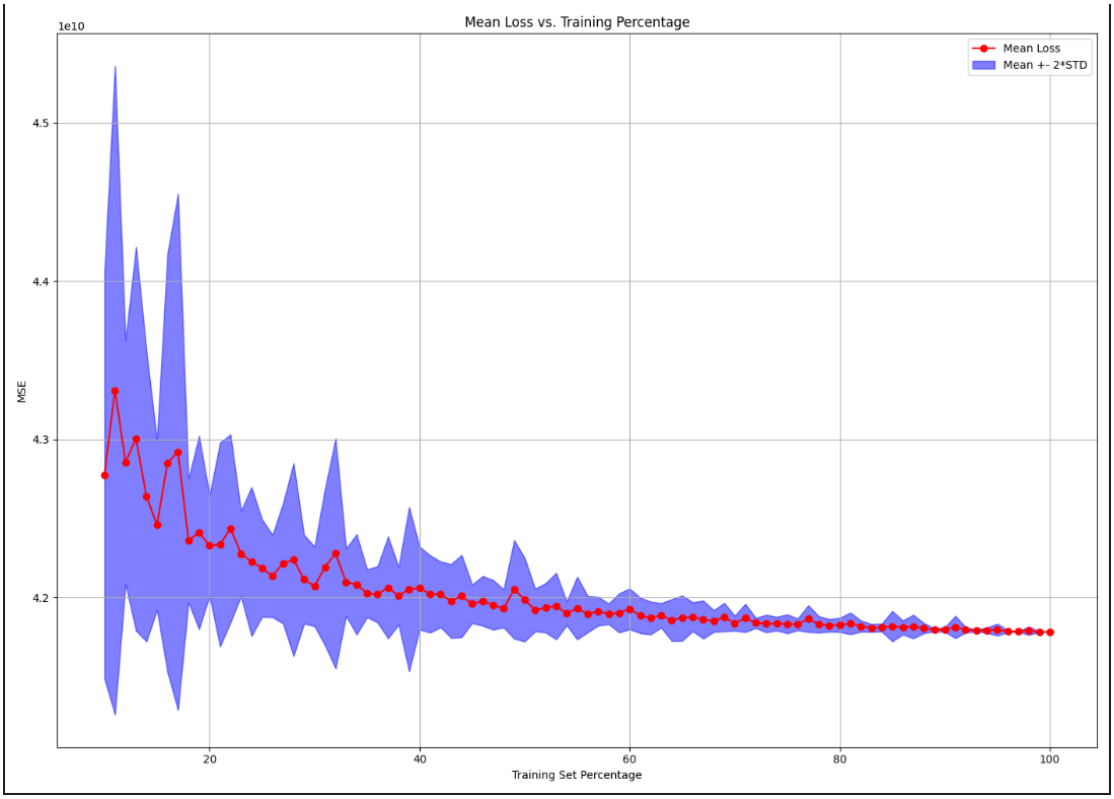
4) גרף 1- שטח אזור בנוי למגורים, ניתן לראות שפיצ'ר זה מאוד רלוונטי למחיר, ציון קורלציית הפירסון שלו הוא 0.7 מה שמעיד על קשר לינארי גבוה וניתן לראות גם בגרף את מגמת העלייה במחיר בהתאם לעליית בשטח הבנוי למגורים. כלומר פיצ'ר זה מאוד רלוונטי למודל.



גרף 2- הפרש בין שטח הבית הכולל לבין שטח הבית הכולל של 15 השכנים הקרובים ביותר אליו. זהו אחד הפיצ'רים שאני הוספתי ולהפתעתי ניתן לראות שהקשר הלינארי מאוד נמוך כלומר אין השפעה של אמיתית על מחיר הבית של ההפרש בין הנתונים הללו. ניתן לראות זאת כי ציון הפירסון קרוב מאוד ל0 וגם בגרף עצמו ניתן לראות שאין קשר לינארי אמיתי ולכן כנראה שאנשים לא מאוד מושפעים מהשוואת שטח ביתם לשטח של בית שכניהם עד כדי עליית מחירים.



(6

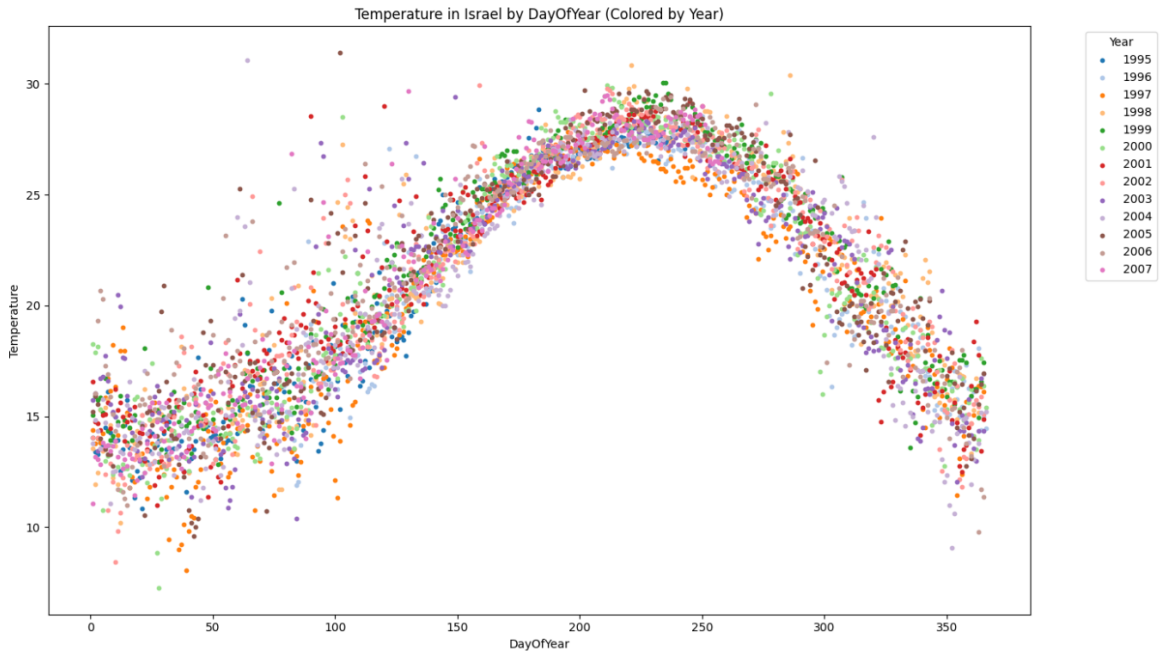


ניתן לראות שMSE נע בין 4.5 לle10 4.2 ומאחר והוא מחושב בצורה ריבועית הloss הוא בעצם ב אזור 200,000 שזה נשמע טווח סביר בהתאם למחירי בתים(מליונים).

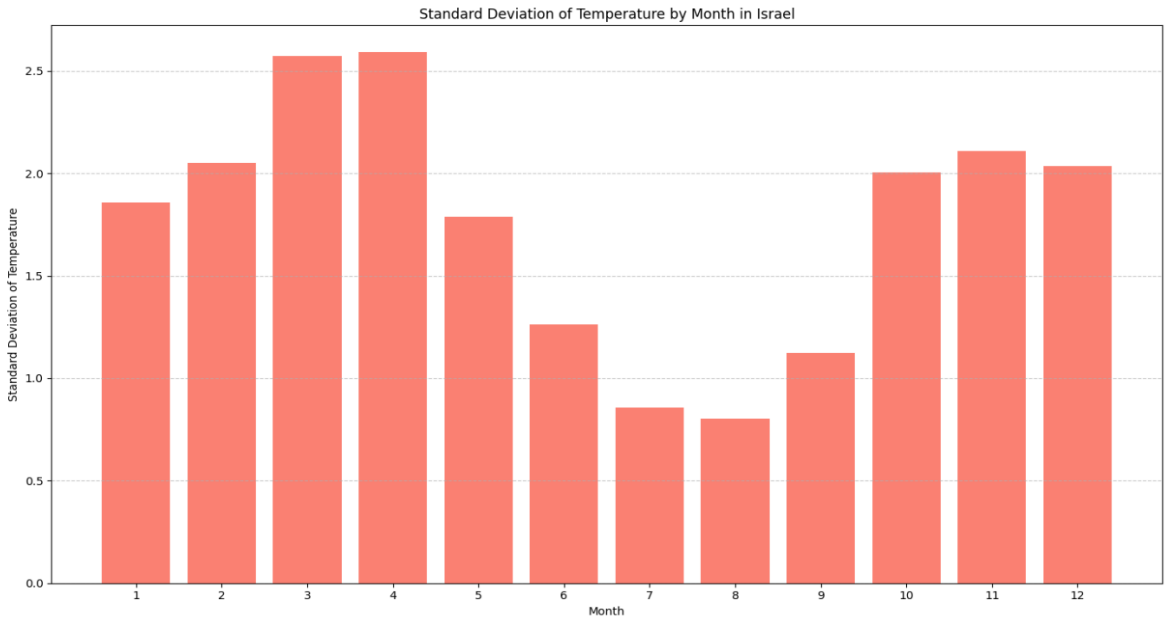
בנוסף ניתן לראות שיש קורלציה בין אחוזי הדגימה לירידה בMSE הכללי, כלומר המודל משתפר ככל שהוא מקבל דגימה יותר גדולה מהנתונים, בהתחלה השינויים חדים אך מאזור ה50 אחוז דגימה הMSE כמעט ולא משתנה כלומר אין לכמות הנתונים שגדלה משמעות בהקטנת הLOSS. לגבי ה2-+ (Confidence interval of mean(loss ניתן לראות שבהתחלה הוא בטווח רחב מאוד מהMSE עצמו אך ככל שאחוז הדגימה עולה הטווח שלו מצטמצם והוא נהיה כמעט זהה לMSE האמיתי כלומר טווח הLOSS של דגימה מתקרב לMSE הממוצע שקיבלנו ככל שאחוז הדגימות עולה, זה מעיד על מודל שמשתפר ככל שהוא לומד יותר.

Polynomial Fitting:

(3

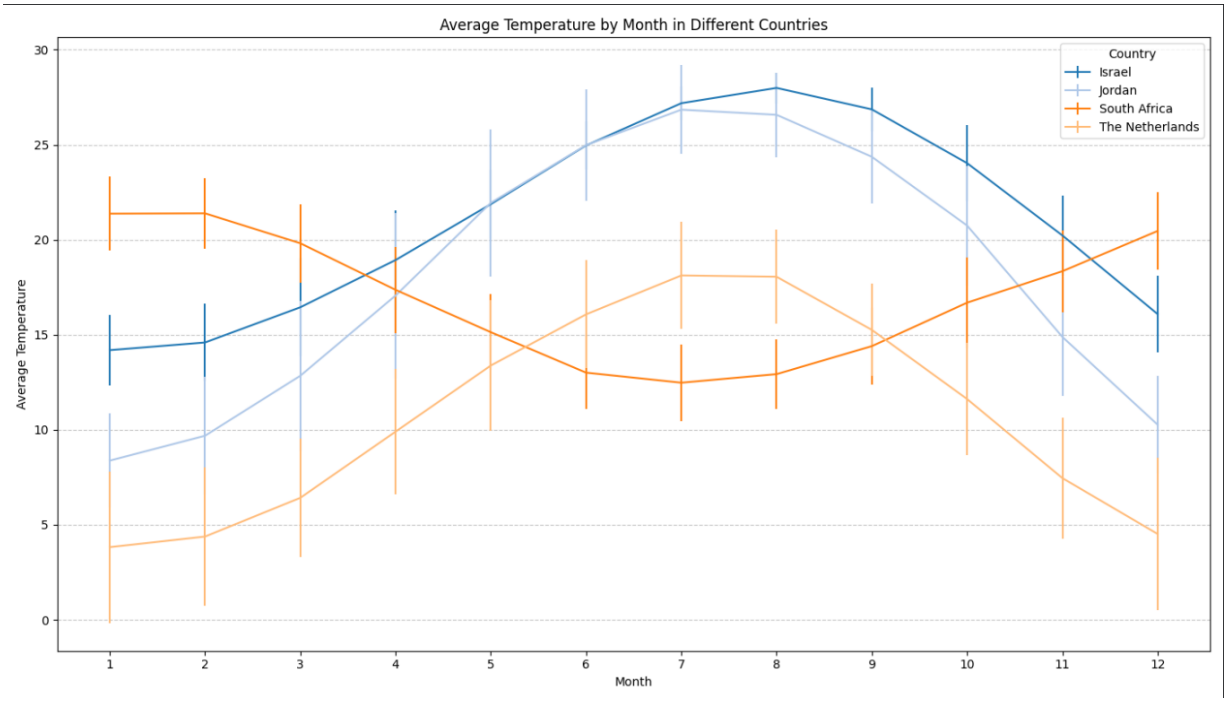


לפי צורת הגרף ניתן לראות שהגרף די מתאים לכל פולינום מדרגה אי זוגית בשל צורתו ולכן אשער שהדרגה המתאימה ביותר היא 3 או 5.



בהתאם לגרף נשים לב שהשונות של טמפרטורה יומית באותו חודש משתנה בין החודשים ולכן נשער שמודל החיזוי יעבוד בצורה הטובה ביותר על חודשים בהם השונות הזאת היא הקטנה ביותר(יולי, אוגוסט) לעומת חודשים בהם השונות היא גדולה (מרץ אפריל) ולכן המודל יצליח במידה שונה לכל חודש.

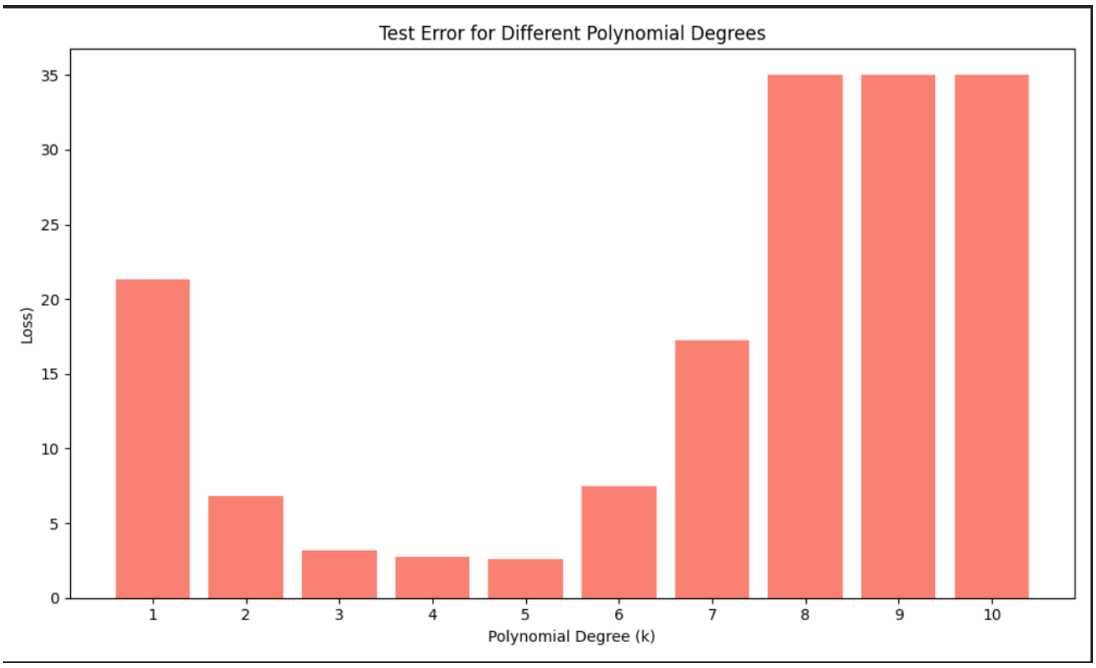
(4



בהתאם לגרף ניתן לראות שישראל ירדן והולנד נצמדות לאותה תבנית אך הולנד עם טמפרטורות כלליות יותר נמוכות לאורך כל השנה. לעומתן דרום אפריקה נצמדת לתבנית הפוכה לגמרי דבר שמסתדר עם העונות ההפוכות בשל הימצאותן בחצי כדור הארץ הדרומי. מפאת תוצאות אלה נסיק שהמודל שיתאים לישראל יתאים כנראה בצורה המקסימלית גם לירדן בשל הדמיון המקסימלי בין הגרפים של שתי המדינות.

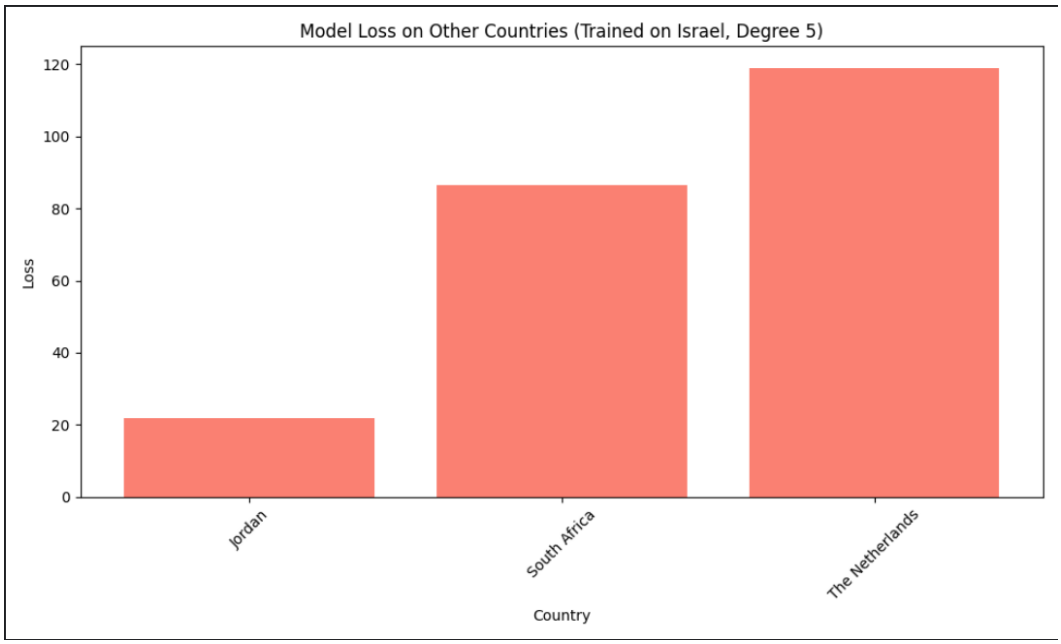
(5

```
k=1, loss=21.320242021509923
k=2, loss=6.825831621422252
k=3, loss=3.171923340889096
k=4, loss=2.7135381423401213
k=5, loss=2.619784398159097
k=6, loss=7.454277989377751
k=7, loss=17.273324151668007
k=8, loss=35.00129019730737
k=9, loss=35.000859274784695
k=10, loss=35.00079890448881
```



ניתן לראות לפי התוצאות שהערך המינימלי של הLOSS מגיע כאשר K=5 כלומר פולינום מדרגה 5 הוא אופטימלי למזער את הLOSS ניתן בנוסף לראות שגם 3 ו4 נותנים ערכים טובים שאולי שווה ליצור מהם מודל פולינומיאלי אך לא כמו 5.

(6



ניתן לראות שבהתאם להשערתנו בשאלה 4 הLOSS הנמוך ביותר לK האופטימלי(5) ירדן היא המדינה שהמודל שפיתחנו הכי מתאים לה כי בה ערך הLOSS הנמוך ביותר שזו השאיפה במודל. מה שכן מפתיע הוא שדרום אפריקה במקום השני על אף שבשאלה 4 צפינו שהולנד תהיה במקום השני כנראה בשל השוני הגדול בטמפרטורות הממוצעות.