1. Prove that: $Ker(\mathbf{X}) = Ker(\mathbf{X}^\top \mathbf{X})$

נוכיח בהכלה דו-כיוונית:

אם $v \in Ker X$ אז $(X^TX)v = X^T(Xv) = X^T \cdot \vec{0} = \vec{0}$

בכיוון השני יהא $v \in Ker X^TX$ ונסמן $X = U\Sigma V^T$ פירוק SVD של X.

נשים לב ש $X^TX = (U\Sigma V^T)^T(U\Sigma V^T) =$

$$= V\Sigma^T U^T U \Sigma V^T = V\Sigma^2 V^T$$

לכן $\Sigma^2 V^T v = \vec{0} \Leftarrow V\Sigma^2 V^T v = \vec{0}$

$V^T v \in Ker \Sigma^2 \overset{*}{=} Ker \Sigma \Leftarrow$

לכן $U^T X v = U^T U \Sigma V^T v = \Sigma V^T v = \vec{0}$

לכן $Xv = U \cdot \vec{0} = \vec{0}$ כלומר $v \in Ker X$ כנדרש.

■

(*) $Ker \Sigma^2 = Ker \Sigma$ כי $v \in Ker \Sigma \iff \forall i: \sigma_i \cdot v_i = 0$

$\iff \forall i: \sigma_i^2 v_i = 0$

$\iff v \in Ker \Sigma^2$ ■

2. Prove that for a square $\overset{d \times d}{\text{matrix}}$ A: $Im\left(A^{\top}\right) = Ker\left(A\right)^{\perp}$

$\subseteq$ אם $v \in Im(A^T)$ אז $\exists$ $x \in \mathbb{R}^d$

כך ש $v = A^T x$.

וכן אם $y \in Ker A$ אז $Ay = 0$

ולכן $\langle v, y \rangle = \langle A^T x, y \rangle = \langle x, Ay \rangle =$

$= \langle x, 0 \rangle = 0$

כלומר $v \in Ker(A)^{\perp}$.

אך ממשפט המימדים $\dim Im A + \dim Ker(A) = d$

$\Leftarrow$ $\dim Im A^T = \dim Im A = d - \dim Ker(A)$

אז $\dim Ker(A)^{\perp} = d - \dim Ker(A) = \dim Im A^T$

ולכן יחד עם ההכלה נסיק $Im A^T = Ker(A)^{\perp}$

$\blacksquare$

3. Let $\mathbf{y} = \mathbf{X}\mathbf{w}$ be a non-homogeneous system of linear equations. Assume that $\mathbf{X}$ is square and not invertible. Show that the system has ∞ solutions $\Leftrightarrow y \perp Ker\left(\mathbf{X}^{\top}\right)$.

$\Leftarrow$ יש ∞ w כך ש- $Xw=y$, בפרט יש w אחד.

כעת יהי $v\in Ker X^T$ אז

$$\langle y, v\rangle = \langle Xw, v\rangle = \langle w, X^T v\rangle = \langle w, 0\rangle = 0$$

כלומר $y \perp Ker(X^T)$.

$\Rightarrow$ $y \perp Ker(X^T)$ כלומר $y \in (Ker(X^T))^\perp$ אבל לפי (2)

$Ker(X^T)^\perp = Im((X^T)^T) = Im(X)$ לכן יש $w_0\in\mathbb{R}^d$ כך ש- $Xw=y$.

בנוסף X אינה הפיכה לכן יש ∞ פתרונות ל- $Xv=0$

נסמן $\{v_n\}_{n=1}^{\infty}$. כעת: $X(w_0+v_n) = Xw_0 + Xv_n = y$ כלומר

$\{w_0+v_n\}_{n=1}^{\infty}$ מהווים ∞ פתרונות ל- $Xw=y$ כנדרש.

4. Consider the (normal) linear system $\mathbf{X}^\top \mathbf{X}\mathbf{w} = \mathbf{X}^\top \mathbf{y}$. Using what you have proved above prove that the normal equations can only have a unique solution (if $\mathbf{X}^\top \mathbf{X}$ is invertible) or infinitely many solutions (otherwise).

אם X^TX הפיכה אז $w=(X^TX)^{-1}X^Ty$ וזהו

פתרון יחיד.

אחרת $X^Ty \perp Ker((X^TX)^T)$ כי אם $v \in Ker(X^TX)=Ker(X)$

מתקיים $\langle X^Ty, v\rangle = \langle y, (X^T)^Tv\rangle = \langle y, Xv\rangle = \langle y, 0\rangle = 0$

קיבלנו X^TX ריבועית ולא הפיכה ולכן משאלה (3) יש

∞ פתרונות למשוואה כנדרש. $\blacksquare$

$(A+B)y = Ax + Bx$

5. Based on Recitation 1 In this question you will prove some properties of orthogonal projection matrices seen in recitation 1. Let $V \subseteq \mathbb{R}^d$, $dim(V) = k$ and let $\mathbf{v}_1, \dots, \mathbf{v}_k$ be an orthonormal basis of V. Define the orthogonal projection matrix $P = \sum_{i=1}^{k} \mathbf{v}_i \mathbf{v}_i^\top$ (notice this is an outer product).

(a) $$P^T = \left[\sum_{i=1}^{k} v_i v_i^T\right]^T = \sum_{i=1}^{k} (v_i v_i^T)^T = \sum_{i=1}^{k} (v_i^T)^T v_i^T = \sum_{i=1}^{k} v_i v_i^T = P$$

(b) נסמן $v_{k+1}, \dots, v_d$ בסיס של $V^\perp$ אז

$$Pv_n = \left(\sum_{i=1}^{k} v_i v_i^T\right) \cdot v_n = \sum_{i=1}^{k} (v_i v_i^T v_n = \sum_{i=1}^{k} v_i \cdot \langle v_i, v_n \rangle = \begin{cases} v_n & n \le k \\ 0 & n > k \end{cases}$$

כאשר אם $n > k$ $\langle v_i, v_n \rangle = 0$ לכל $i \in [k]$ כי $v_n \in V^\perp$

ואם $n \le k$ $\langle v_i, v_n \rangle = 1 \iff i = n$ כי $v_1, \dots, v_k$

בסיס אורתונורמלי.

כלומר מצאנו ש $v_1, \dots, v_k$ וקטורים עצמיים עם ע"ע 1

״ $v_{k+1}, \dots, v_d$ ״ ״ ״ ״ 0

ואלו כל הערכים העצמיים משיקולי מימדים של המרחבים העצמיים.

$\blacksquare$

(c) יהא $v\in V$ אז $v=\sum_{j=1}^{k} a_j\cdot v_j$ לכן

$$Pv=\left(\sum_{i=1}^{k} v_i v_i^T\right)v=\sum_{i=1}^{k} v_i v_i^T\left(\sum_{j=1}^{k} a_j v_j\right)=\sum_{i=1}^{k}\sum_{j=1}^{k} v_i v_i^T a_j v_j=$$

$$=\sum_{i=1}^{k}\sum_{j=1}^{k} a_j v_i\langle v_i,v_j\rangle=\sum_{i=1}^{k} a_i v_i=v$$

כאשר המעבר האחרון נובע מכך ש- $v_1,\dots,v_k$ בסיס א"נ

(d) נסמן $P=UDU^T$ פירוק EVD של P (יש כזה
כי P סימטרית) ונקבל $D=\begin{bmatrix}1&&&&\\&1&&&\\&&\ddots&&\\&&&0&\\&&&&0\end{bmatrix}$ כי אלו הע"ע
של P כמובן $D^2=D$ ולכן

$$P^2=(UDU^T)(UDU^T)=UD^2U^T=UDU^T=P$$

(e) נובע ישירות מ-(d):

$$(I-P)P=P-P^2=0$$

2.3 Least Squares

Based on Lecture 2 and Recitation 3 Given a sample $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$, the ERM rule for linear regression w.r.t. the squared loss is

$$\hat{\mathbf{w}} \in \underset{\mathbf{w}\in\mathbb{R}^d}{\operatorname{argmin}} \; ||\mathbf{X}\mathbf{w} - \mathbf{y}||^2$$

where $\mathbf{X}$ is the design matrix of the linear regression with rows as samples and y the vector of responses. Let $\mathbf{X} = U\Sigma V^\top$ be the SVD of $\mathbf{X}$, where U is a $m \times m$ orthonormal matrix, Σ is a $m \times d$ diagonal matrix, and V is an $d \times d$ orthonormal matrix. Let $\sigma_i = \Sigma_{i,i}$ and note that only the non-zero σ_i-s are singular values of $\mathbf{X}$. Recall that the pseudoinverse of $\mathbf{X}$ is defined by $\mathbf{X}^\dagger = V\Sigma^\dagger U^\top$ where $\Sigma^\dagger$ is an $d \times m$ diagonal matrix, such that

$$\Sigma^\dagger_{i,i} = \begin{cases} \sigma_i^{-1} & \sigma_i \neq 0 \\ 0 & \sigma_i = 0 \end{cases}$$

6. Show that if $\mathbf{X}^\top\mathbf{X}$ is invertible, the general solution we derived in recitation equals to the solution you have seen in class. For this part, assume that $\mathbf{X}^\top\mathbf{X}$ is invertible.

נכתוב $X = U\Sigma V^T$ פירוק SVD ל X אז

$X^TX = V\Sigma^TU^TU\Sigma V^T = V\Sigma^T\Sigma V^T$ ולכן שהיא הפיכה גם

$\Sigma^T\Sigma$ הפיכה ומתקיים $(X^TX)^{-1} = V(\Sigma^T\Sigma)^{-1}V^T$.

לכן $\hat{w} = (X^TX)^{-1}X^Ty = V(\Sigma^T\Sigma)^{-1}V^TV\Sigma^TU^T =$

$$= V(\Sigma^T\Sigma)^{-1}\Sigma^TU^Ty$$

נשים לב ש $X^\dagger = V\Sigma^\dagger U^T$ עם כי ש $\Sigma^\dagger = (\Sigma^T\Sigma)^{-1}\Sigma^T$

ואז נקבל ש $\hat{w} = X^\dagger y$ כנדרש.

נשים לב ש

$$\Sigma^T\Sigma = d\Bigg\{\underbrace{\begin{bmatrix}\sigma_1 & & 0 & 0\\ & \ddots & & \\ 0 & & \sigma_d & 0\end{bmatrix}}_{m}\underbrace{\begin{bmatrix}\sigma_1 & & 0\\ 0 & \ddots & \\ & & \sigma_d\\ & 0 & \end{bmatrix}}_{d}\Bigg\}m = \overbrace{\begin{bmatrix}\sigma_1^2 & & \\ & \ddots & \\ & & \sigma_d^2\end{bmatrix}}^{d}\Bigg\}d$$

ומכאן שהיא הפיכה בפרט $\sigma_i \neq 0$ לכל $i\in[d]$, ומתקיים

$$(\Sigma^T\Sigma)^{-1} = \overbrace{\begin{bmatrix}(\sigma_1)^{-2} & & \\ & \ddots & \\ & & (\sigma_d)^{-2}\end{bmatrix}}^{d}\Bigg\}d$$

לכן

$$(\Sigma^T\Sigma)^{-1}\Sigma^T = d\Bigg\{\overbrace{\begin{bmatrix}(\sigma_1)^{-2} & & \\ & \ddots & \\ & & (\sigma_d)^{-2}\end{bmatrix}}^{d}\underbrace{\begin{bmatrix}\sigma_1 & & 0 & 0\\ & \ddots & & \\ 0 & & \sigma_d & 0\end{bmatrix}}_{m}\Bigg\}d =$$

$$= d\Bigg\{\underbrace{\begin{bmatrix}\sigma_1^{-1} & & 0 & 0\\ & \ddots & & \\ 0 & & \sigma_d^{-1} & 0\end{bmatrix}}_{m} = \Sigma^+$$

כנדרש ∎

7. Show that $\mathbf{X}^\top\mathbf{X}$ is invertible if and only if $\text{span}\{\mathbf{x}_1,\ldots,\mathbf{x}_m\}=\mathbb{R}^d$

⇐ אם X^TX הפיך אז $\dim Ker(X^TX)=0$

לכן $\dim Ker(X)=0$ בפרט X בדרגה מלאה

ואנחנו יודעים ש $d\le m$ לכן $rank(X)=d$ כלומר

לכן $Im X=\mathbb{R}^d$ $\quad Span\{x_1,\ldots,x_m\}=Im X=\mathbb{R}^d$

∎

8. Recall that if $\mathbf{X}^\top\mathbf{X}$ is not invertible then there are many solutions. Show that $\hat{\mathbf{w}}=\mathbf{X}^\dagger\mathbf{y}$ is the solution whose L_2 norm is minimal. That is, show that for any other solution $\overline{\mathbf{w}}$, $||\hat{\mathbf{w}}||\le||\overline{\mathbf{w}}||$.

Hints:

- Recall that the rank of $\mathbf{X}$ and the rank of $\mathbf{X}^\top\mathbf{X}$ are determined by the number of singular values of $\mathbf{X}$. If you are not sure why this is true, go over recitation 1.
- Which coordinates must satisfy $\widehat{w}_i=\overline{w}_i$? What is the value of $\widehat{w}_i$ for the other coordinates? If you are not sure, go back to the derivation of $\widehat{\mathbf{w}}$ (see recitation 4).

אם X^TX לא הפיך אז יש ל $X^TXw=X^Ty$ ∞ פתרונות.

יהי $\bar{w}\in\mathbb{R}^d$ כך ש $X^TX\bar{w}=X^Ty$ נראה ש $\|\hat{w}\|\le\|\bar{w}\|$.

אם ל X יש $k<d$ ערכים סינגולריים שונים מאפס אז

תרגול 4

$\hat{w}_i=\bar{w}_i$ לכל $i\le k$ ולכל $i>k$ $\hat{w}=0$ לכן

$$\|\hat{w}\|=\sum_{i=1}^{d}\hat{w}_i^2=\sum_{i=1}^{k}\hat{w}_i^2+\sum_{i=k+1}^{d}\hat{w}_i^2=\sum_{i=1}^{k}\hat{w}_i^2+0=\sum_{i=1}^{k}\bar{w}_i^2\le\sum_{i=1}^{d}\bar{w}_i^2=\|\bar{w}_i\|$$

∎

Describe in details the analysis process that lead you to the decisions of:

- Which features to keep and which not?
- Which features are categorical how how did you treat them?
- What other features did you design and what is the logic behind creating them?
- How did you treat invalid/missing values?
- Explain any additional processing performed on the data.

The answers to these question should be added to your `Answers.pdf` file.

ראשית הסרתי רכיבי דאטה שנראים לי "מספריים" או חסרי משמעות כמו: id, long, lat, date.

את שאר הרכיבים השארתי ועל פניו נראה שיש קשר בין הנתון עצמו למחיר הנכס.

רכיב ה: zipcode הוא קטגורי וטיפלתי בו דרך המרתו להצגה ע"י חלוקה לעמודות. כדי שלא יווצרו יותר מדי עמודות וגם מכיוון שרוב המידע הקטגורי הרלוונטי נמצא שם, הקטנתי את הספרה האחרונה בכל קוד.

הוספתי מספר עמודות נוספות:

1. יחס בין שטח הסלון לשטח הכולל של הדירה
2. תאריך אחרון ששופץ או נבנה

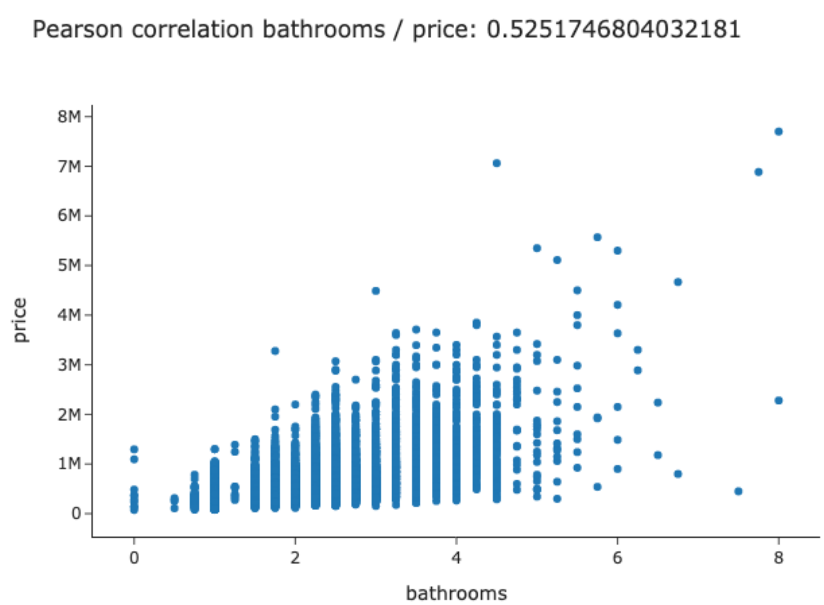
כמו כן הוספתי יחס בין גודל הנכס לגודל של 15 הנכסים

הקורלציה אינה רבה.

ערכים חסרים השלמתי במקרה של דאטה-בדיקה ע"י ממוצע

והסרתי את השורה לגמרי במקרה של דאטה-אימון.

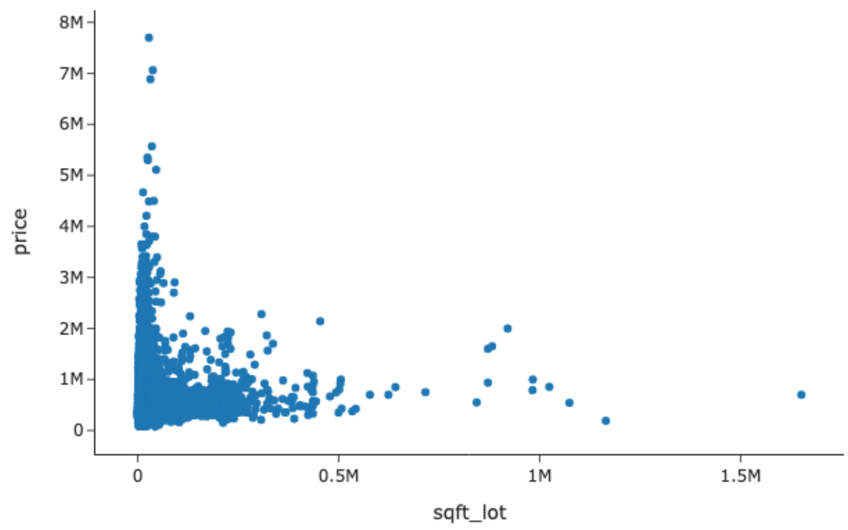
Choose two features, one that seems to be beneficial for the model and one that does not. In your `Answers.pdf` add the graphs of these two chosen features and explain how do you conclude if they are beneficial or not.



ניתן לראות שהמקדם יחסית גבוה 0.52 לכן קיים קשר לינארי

בין מחיר הנכס לכמות חדרי השירותים.

Pearson correlation sqft_lot / price: 0.08982134888822416



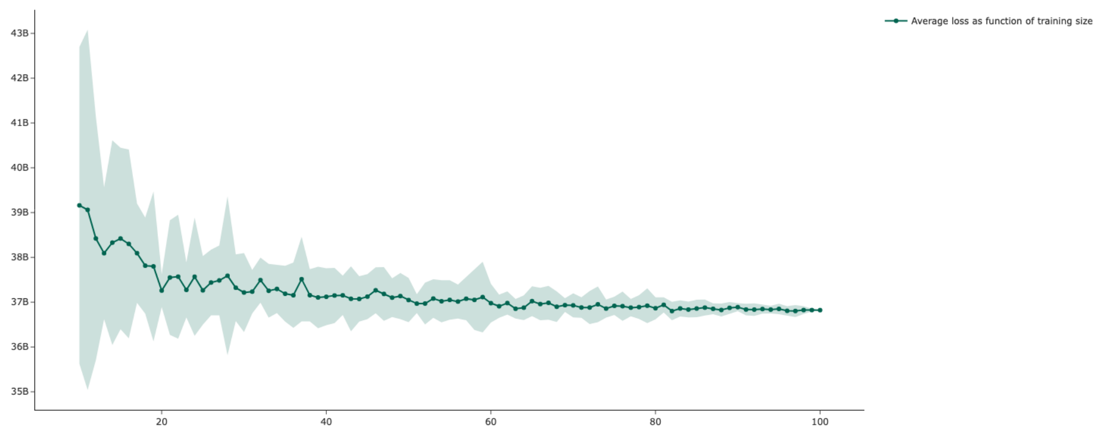
קורלציה (חלק מוצג...) נמוכה זה שלא נראה שיש מגמה

קשר לינארי ראוי לתאר ורואים שהמקדם נמוך יחסית: 0.09.

4. Fit a linear regression model over increasing percentages of the *training set* and measure the loss over the *test set*:
   - Iterate for every percentage $p = 10\%, 11\%, ..., 100\%$ of the training set.
   - Sample $p\%$ of the train set. You can use the `pandas.DataFrame.sample` function.
   - Repeat sampling, fitting and evaluating 10 times for each value of p.

   Plot the mean loss as a function of $p\%$, as well as a confidence interval of $mean(loss) \pm 2 * std(loss)$. If implementing using the `Plotly` library, see how to create the confidence interval in Chapter 2 - Linear Regression code examples.

   Add the plot to the `Answers.pdf` file and explain what is seen. Address both trends in loss and in confidence interval as function of training size. What can we learn about the estimator $\hat{y}_i$ in terms of estimator properties?



ניתן לראות שככל שאנו מגדילים את אחוז גדל יותר את דגימות

אך מתייצב

ה: loss יורד ומגיע MSE גם השונות קטנה.

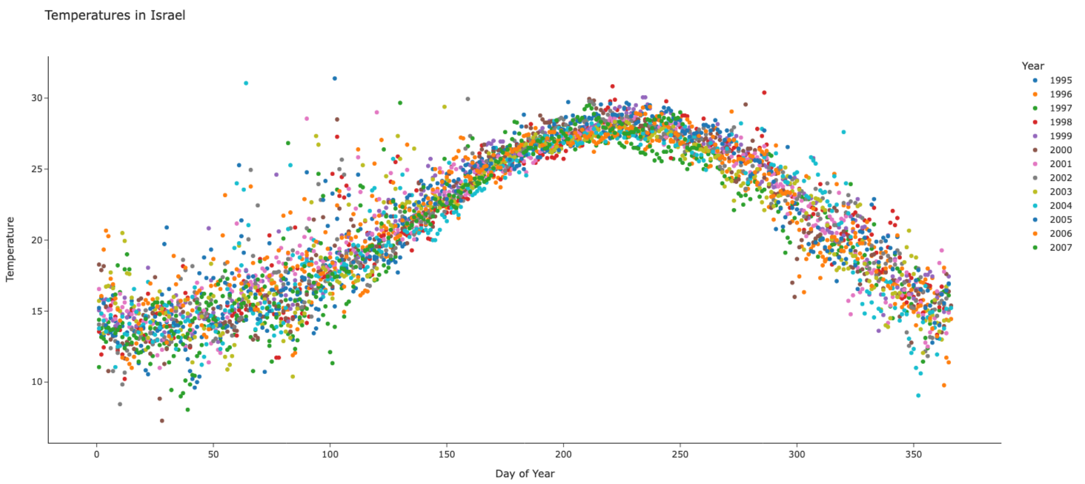
ה: bias יורד כי עם ריבוי שונות של דגימות קיבלנו תוצאות

יותר מגוונות ושונות נשען על שפע. ועם השונות קטנה כי אין

הבדל דרמתי הדגימות רק עשה הרבה.

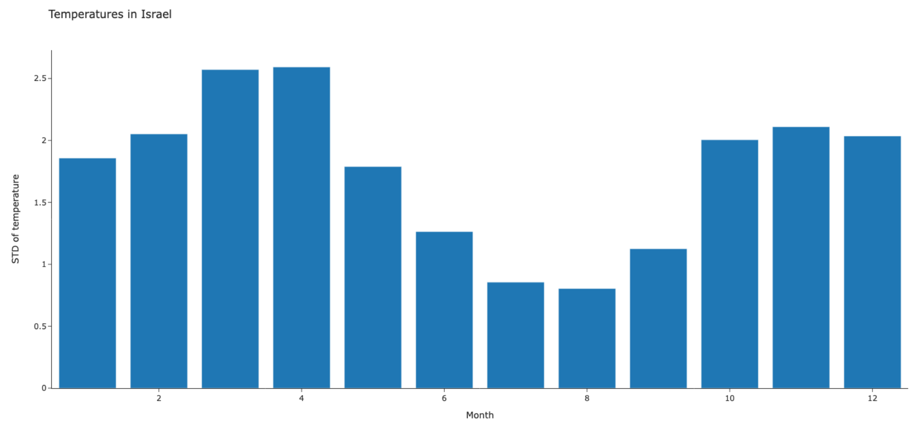
2. Subset the dataset to caintain samples only from the country of Israel. Investigate how the average daily temperature (‘Temp‘ column) change as a function of the ‘DayOfYear‘.
   - Plot a scatter plot showing this relation, and color code the dots by the different years (make sure color scale is discrete and not continuous). What polynomial degree might be suitable for this data?
   - Group the samples by ‘Month‘ (have a look at the pandas ‘groupby‘ and ‘agg‘ functions) and plot a bar plot showing for each month the standard deviation of the daily temperatures. Suppose you fit a polynomial model (with the correct degree) over data sampled uniformly at random from this dataset, and then use it to predict temperatures from random days across the year. Based on this graph, do you expect a model to succeed equally over all months or are there times of the year where it will perform better than on others? Explain your answer.

   Add both plots and answers to the `Answers.pdf` file.



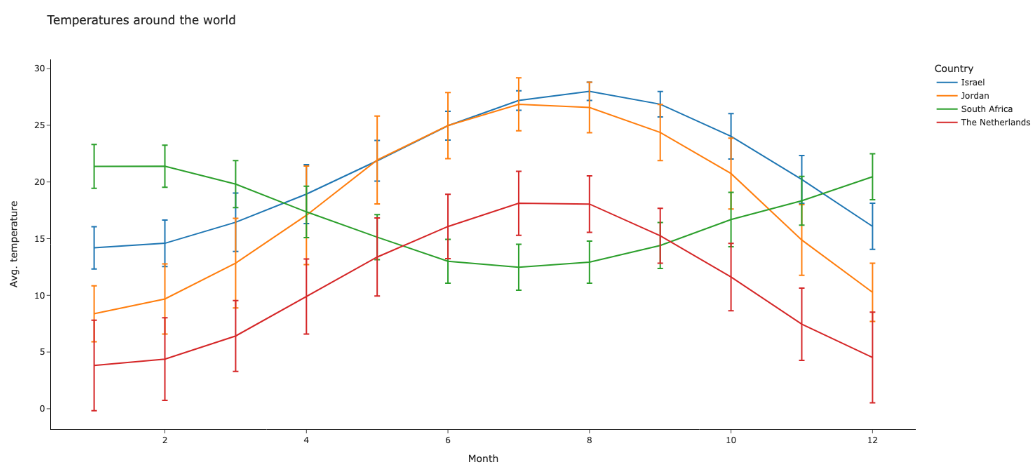
ניתן לראות שתי נקודות קיצון אך נצפה שגיאות תתאבס

לפחות שתי נקודות כלומר פולינום ממעלה 3 לפחות:



ניתן לשער שהמודל יתאים הכי טוב ליולי-אוגוסט שם השונות נמוכה יותר.

3. Returning to the full dataset, group the samples according to ‘`Country`‘ and ‘`Month`‘ and calculate the average and standard deviation of the temperature. Plot a line plot of the average monthly temperature, with error bars (using the standard deviation) color coded by the country. If using `plotly.express.line` have a look at the `error_y` argument.

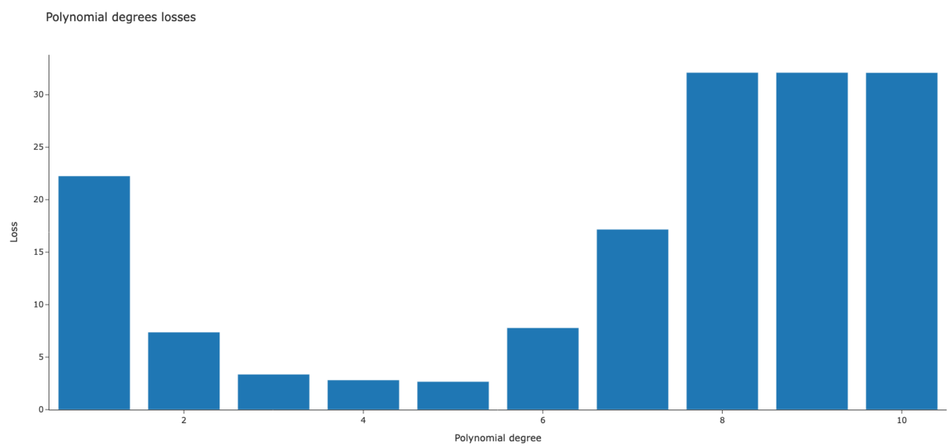


ניתן לראות שישראל וירדן מתנהגים מאוד דומה, הולנד מתנהגת דומה אך עם intercept שונה משמעותית ודרום אפריקה מתנהגת בצורה שונה אשר מגיעה אל עוד הכי טוב אל ירדן.

4. Over the subset containing observations only from Israel perform the following:
   - Randomly split the dataset into a training set (75%) and test set (25%).
   - For every value $k \in [1, 10]$, fit a polynomial model of degree k using the training set.
   - Record the loss of the model over the test set, rounded to 2 decimal places.

   Print the test error recorded for each value of k. In addition plot a bar plot showing the test error recorded for each value of k. Based on these which value of k best fits the data? In the case of multiple values of k achieving the same loss select the simplest model of them. Are there any other values that could be considered?

```
[22.26, 7.38, 3.37, 2.82, 2.67, 7.8, 17.18, 32.11, 32.11, 32.1]
```

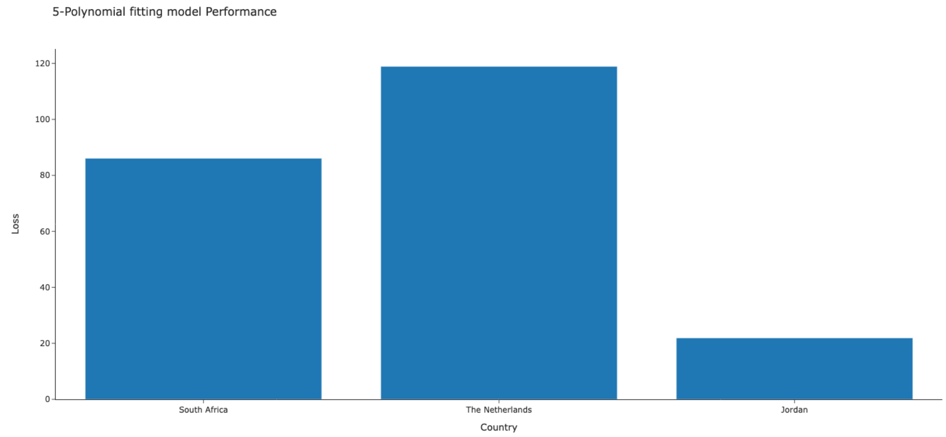


k=5 בעלת ה- loss הנמוך ביותר.

3,4 בעלות loss דומה ל-5 ולכן ניתן לשקול לאחד המשוערים

ולבחור באחד מהם.

5. Fit a model over the entire subset of records from Israel using the *k* chosen above. Plot a bar plot showing the model's error over each of the other countries. Explain your results based on this plot and the results seen in question 3.



כפי שדיברנו, הולנד ודרום אפריקה שאינן דומות לישראל

לא קיבלו תוצאה טובה כמו ירדן יחסית קרובה.