

Theoretical Part

2.1 Mathematical Background

2.1.1 Linear Algebra

1. • first, calculate $A^T A$:

$$\begin{matrix} 3 \times 2 & 2 \times 3 \end{matrix} \quad \begin{bmatrix} 1 & 1 \\ 1 & -1 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \\ 1 & -1 & 2 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 2 \\ 0 & 2 & -2 \\ 2 & -2 & 4 \end{bmatrix}$$

• find the eigenvalues and eigenvectors by the characteristic polynom of $A^T A$:

$$P_{A^T A}(\lambda) = |A^T A - \lambda I| = \begin{vmatrix} 2-\lambda & 0 & 2 \\ 0 & 2-\lambda & -2 \\ 2 & -2 & 4-\lambda \end{vmatrix}$$
$$= (2-\lambda) \begin{vmatrix} 2-\lambda & -2 \\ -2 & 4-\lambda \end{vmatrix} + 2 \begin{vmatrix} 0 & 2 \\ 2-\lambda & -2 \end{vmatrix}$$

$$= (2-\lambda)[(2-\lambda)(4-\lambda) - 4] - 4(2-\lambda)$$
$$= (2-\lambda)[4-6\lambda+\lambda^2] - 4(2-\lambda)$$

$$\begin{aligned}
 &= (2-\lambda)(4-6\lambda+\lambda^2-4) = (2-\lambda)(\lambda^2-6\lambda) \\
 &= (2-\lambda)(\lambda-6)\lambda
 \end{aligned}$$

The eigenvalues are $\lambda_1 = 0$, $\lambda_2 = 6$, $\lambda_3 = 2$

- will find the eigenvector for $\lambda_1 = 0$:

$$(A^T A - 0I)v = 0$$

$$\left[\begin{array}{ccc} 2 & 0 & 2 \\ 0 & 2 & -2 \\ 2 & -2 & 4 \end{array} \right] \xrightarrow{R_3 \rightarrow R_3 - R_1} \left[\begin{array}{ccc} 2 & 0 & 2 \\ 0 & 2 & -2 \\ 0 & -2 & 2 \end{array} \right] \xrightarrow{R_3 \rightarrow R_3 + R_2} \left[\begin{array}{ccc} 2 & 0 & 2 \\ 0 & 2 & -2 \\ 0 & 0 & 0 \end{array} \right]$$

$$\begin{aligned}
 R_1 \rightarrow R_1 \cdot \frac{1}{2} &\quad \left[\begin{array}{ccc} 1 & 0 & 1 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{array} \right] & x_3 = t_1, \quad x_2 = t_1, \quad x_1 = -t_1
 \end{aligned}$$

$$\text{The solutions set is: } \left\{ \begin{bmatrix} -t_1 \\ t_1 \\ t_1 \end{bmatrix} \mid t_1 \in \mathbb{R} \right\} = \left\{ \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix} \right\}$$

- will find the eigenvector for $\lambda_1 = 2$:

$$\left[\begin{array}{ccc} 0 & 0 & 2 \\ 0 & 0 & -2 \\ 2 & -2 & 2 \end{array} \right] \xrightarrow{R_2 \rightarrow R_2 + R_1} \left[\begin{array}{ccc} 0 & 0 & 2 \\ 0 & 0 & 0 \\ 2 & -2 & 2 \end{array} \right] \xrightarrow{R_1 \leftrightarrow R_3} \left[\begin{array}{ccc} 2 & -2 & 2 \\ 0 & 0 & 0 \\ 0 & 0 & 2 \end{array} \right]$$

$$R_1 \rightarrow R_1 - R_3 \left[\begin{array}{ccc} 2 & -2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 2 \end{array} \right] R_2 \leftrightarrow R_3 \left[\begin{array}{ccc} 1 & -1 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{array} \right] R_2 \rightarrow R_2 \cdot \frac{1}{2} \left[\begin{array}{ccc} 1 & -1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{array} \right]$$

$$x_2 = t_1, x_3 = 0, x_1 = t_1$$

The solution set is: $\left\{ \begin{bmatrix} t_1 \\ t_1 \\ 0 \end{bmatrix} \mid t_1 \in \mathbb{R} \right\} = \left\{ \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \right\}$

- will find the eigenvector for $\lambda_1 = 6$:

$$\left| \begin{array}{ccc} -4 & 0 & 2 \\ 0 & -4 & -2 \\ 2 & -2 & -2 \end{array} \right| R_1 \rightarrow R_1 + 2R_3 \left| \begin{array}{ccc} 0 & -4 & -2 \\ 0 & -4 & -2 \\ 2 & -2 & -2 \end{array} \right| R_2 \rightarrow R_2 - R_1 \left| \begin{array}{ccc} 0 & -4 & -2 \\ 0 & 0 & 0 \\ 2 & -2 & -2 \end{array} \right|$$

$$R_1 \leftrightarrow R_3 \left[\begin{array}{ccc} 2 & -2 & -2 \\ 0 & 0 & 0 \\ 0 & -4 & -2 \end{array} \right] R_2 \leftrightarrow R_3 \left[\begin{array}{ccc} 2 & -2 & -2 \\ 0 & -4 & -2 \\ 0 & 0 & 0 \end{array} \right] R_1 \rightarrow R_1 - \frac{1}{2}R_2 \left[\begin{array}{ccc} 2 & 0 & -1 \\ 0 & -4 & -2 \\ 0 & 0 & 0 \end{array} \right]$$

$$R_2 \rightarrow R_2 \cdot \frac{1}{4}$$

$$R_1 \rightarrow R_1 \cdot \frac{1}{2} \left[\begin{array}{ccc} 1 & 0 & -\frac{1}{2} \\ 0 & 1 & \frac{1}{2} \\ 0 & 0 & 0 \end{array} \right]$$

$$x_3 = t_1, x_2 = -\frac{t_1}{2}, x_1 = \frac{t_1}{2}$$

The solution set: $\left\{ t_1 \begin{bmatrix} -\frac{1}{2} \\ -\frac{1}{2} \\ 1 \end{bmatrix} \right\}$

$$= \left\{ \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix} \right\}$$

- normalize the eigenvectors:

$$\hat{V}_0 = \begin{bmatrix} -\frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{3}} \end{bmatrix}, \quad \hat{V}_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{bmatrix}, \quad \hat{V}_6 = \begin{bmatrix} \frac{1}{\sqrt{6}} \\ -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \end{bmatrix}$$

- Single values are: $0, \sqrt{2}, \sqrt{6}$
- right single values are:

$$V = \begin{bmatrix} -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & \frac{\sqrt{2}}{\sqrt{6}} \end{bmatrix} \quad \Sigma = \begin{bmatrix} \sqrt{6} & 0 & 0 \\ 0 & \sqrt{2} & 0 \end{bmatrix}$$

- find U by computing $AV = U\Sigma^{-1}$:

$$\begin{matrix} 2 \times 3 & 3 \times 3 & 2 \times 2 & 2 \times 3 \end{matrix} \quad \begin{bmatrix} 1 & 1 & 0 \\ 1 & -1 & -2 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} \\ \frac{\sqrt{2}}{\sqrt{6}} & 0 & \frac{1}{\sqrt{3}} \end{bmatrix} = \begin{bmatrix} x_1 & x_2 \\ x_3 & x_4 \end{bmatrix} \begin{bmatrix} \sqrt{6} & 0 & 0 \\ 0 & \sqrt{2} & 0 \end{bmatrix}$$

$$\text{so, } x_1 = 0, x_2 = 1$$

$$x_3 = 1, x_4 = 0$$

Then:

$$\begin{bmatrix} 0 & \sqrt{2} & 0 \\ \sqrt{6} & 0 & -\frac{2}{\sqrt{3}} \end{bmatrix} = \begin{bmatrix} x_1\sqrt{6} & x_2\sqrt{2} & 0 \\ x_3\sqrt{6} & x_4\sqrt{2} & 0 \end{bmatrix}$$

$$U = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

2.1.00

לט' $V \oplus U$ נסatisfי הדרישות הנקויה $\mathbb{R}^m \ni u \in V$ ו $\mathbb{R}^n \ni v \in U$ מתקיים $u+v \in V \oplus U$

$$A = \begin{bmatrix} v_1 u_1 & \dots & v_1 u_m \\ v_2 u_1 & \ddots & v_2 u_m \\ \vdots & & \vdots \\ v_n u_1 & & v_n u_m \end{bmatrix}$$

A הוא מטריצת $n \times m$ ו A^T הוא מטריצת $m \times n$.

A הוא מטריצת $n \times m$. $\text{Rank}(A) = \text{Rank}(A^T)$, A^T הוא מטריצת $m \times n$.

$$A^T = \begin{bmatrix} v_1 u_1 & \dots & v_n u_1 \\ v_1 u_2 & \ddots & v_n u_2 \\ \vdots & & \vdots \\ v_1 u_m & & v_n u_m \end{bmatrix}$$

בנוסף למשהו שקיים ב- A (ב- A^T לא), A מתקיים גם ב- A^T .

הו שקיים ב- A^T לא ב- A נקרא $\text{rank}(A^T) - \text{rank}(A)$.

ההוכחה מושגת באמצעות אמצעים של חישוב דרגה. נניח R_1, R_2, \dots, R_n מטריצות שמקיימות $R_2 \rightarrow R_2 - R_1 \cdot \frac{u_2}{u_1}$, $R_i \rightarrow R_i - R_1 \cdot \frac{u_i}{u_1}$ ו $R_i \rightarrow R_i - R_2 \cdot \frac{u_i}{u_2}$.

בנוסף לכך, ניקח v_1, v_2, \dots, v_n מטריצות שמקיימות $v_2 \rightarrow v_2 - v_1 \cdot \frac{u_2}{u_1}$, $v_i \rightarrow v_i - v_1 \cdot \frac{u_i}{u_1}$ ו $v_i \rightarrow v_i - v_2 \cdot \frac{u_i}{u_2}$.

$$\text{rank}(A^T) = \text{rank}(A) - e$$

3 p'go

ר' פון ניירמן, כרך ג'. $x = a_1u_1 + a_2u_2 + \dots + a_nu_n - e$ ר' פון ניירמן נגזרת

$$\langle x, u_i \rangle = \langle a_1 u_1 + \dots + a_n u_n, u_i \rangle$$

$$= \alpha_1 \langle u_1, u_i \rangle + \dots + \alpha_i \langle u_i, u_i \rangle + \dots + \alpha_n \langle u_n, u_i \rangle = \overset{*}{\alpha_i} \langle u_i, u_i \rangle$$

$O = \langle u_k, u_j \rangle \Leftrightarrow u_k \perp u_j$ גורם * . עפיה נסביהה בינהו *

ב $k = j$ ג'לה ר"ג נתקנה, כוונתנו היא $(u_1, \dots, u_n) - e$ נכל

$\Rightarrow \mathbf{1} = \langle \mathbf{u}_r, \mathbf{u}_r \rangle$, r សម្រាប់ ពី $1 \leq k < j \leq n - e$ $\Rightarrow \mathbf{u}_k \perp \mathbf{u}_j$ $k \neq j$ -e

$$a_i = \langle x, u_i \rangle - \ell \quad \text{for } j, \text{ if } 1 \leq r \leq n - \ell$$

.4 foo

$$(V_i, V_i^T)^T = (V_i^T)^T V_i^T = V_i V_i^T \text{ איפרנסונסיל } : P^T = P \quad P^T = P \text{ נסוי } \quad (a)$$

כפינט $V_i V_i^T$ אונגרית. פיקט $V_i V_i^T$ אונגרית. הינה $\sum_{i=1}^p V_i V_i^T = e$ יפה.

$P(v_j) = v_j : P \cap N(v_1, \dots, v_k) \ni v_j$ סביר מכך (b)

$$P(v_j) = \sum_i^d v_i v_i^T \cdot v_j = \sum_i^d v_i < v_i^T, v_j > = v_j < v_j^T, v_j > = v_j$$

$$O = \langle v_k, v_j \rangle, k \neq j \in \{1, 2, \dots, n\} \subset O(O(v_1, \dots, v_k) - e) \subset$$

נניח P הוא פולינומיאלי ב- λ ו- e שקיים $\lambda = \langle v_k, v_k \rangle$!

$\lambda = \lambda - e$ רצוי נסsat מ- λ ב- e ו- $v_k = (v_1, \dots, v_k) - e$

נניח $P(u) = \lambda u - e$ ו- $0 + u$ נורמלית $1 + \lambda$ "ב" פולינומיאלי נסsat ע"י

הנורמליזציה של v_k יתנו $v_k = (v_1, \dots, v_k, u)$ נורמליזציה נסsat

בנוסף לה נורמליזציה של v_k נסsat. אופרציונל $(v_1, \dots, v_k) - e$ מושג

$P - 1$ "ב" פולינומיאלי ב- e ו- v_k נסsat. $k = \dim(V)$

$V = \alpha_1 v_1 + \dots + \alpha_k v_k$ אופרציונל נסsat. נסsat v נסsat. (c)

נסsat נסsat

$$P(v) = P(\alpha_1 v_1 + \dots + \alpha_k v_k) = \alpha_1 P(v_1) + \dots + \alpha_k P(v_k)$$

b ת'זון *
 $= \alpha_1 v_1 + \dots + \alpha_k v_k = v$

$$P \cdot P = \left(\sum_{i=1}^k v_i v_i^\top \right) \left(\sum_{j=1}^k v_j v_j^\top \right) = \sum_{i=1}^k \sum_{j=1}^k v_i v_i^\top \cdot v_j v_j^\top$$

$$* = \sum_{i=1}^k v_i v_i^\top = P$$

$$1 = \langle v_j^\top, v_j \rangle \quad j=i \text{ נסsat} \quad 0 = \langle v_i^\top, v_j \rangle \quad i \neq j \text{ נסsat *} \quad (d)$$

d ת'זון

$$(I - P)P = IP - P^2 = P - P^2 \xrightarrow{\uparrow} P - P = 0 \quad (e)$$

2.1.2 Multivariable calculus

: ר'גון 'מ' $s(\sigma) = f(\sigma) - y$, $g(\sigma) = \frac{1}{2} \|\sigma\|^2$. 5
 $(g \circ s)(\sigma)$ הינו מינימום רגולרי

$$\begin{aligned} J_\sigma(h) &= J_\sigma(g \circ s) = J_{s(\sigma)}(g) \cdot J_\sigma(s) = \sigma^\top \cdot J_\sigma(s) \\ &= s(\sigma)^\top \cdot J_\sigma(s) = (f(\sigma) - y)^\top \cdot J_\sigma(s) \\ &= (f(\sigma) - y)^\top \cdot \nabla f \end{aligned}$$

$$\nabla_\sigma(h) = J_\sigma(h)^\top = J_\sigma(s)^\top (f(\sigma) - y)$$

: ר'גון $j=i$ דמיון ליניארי . 6

$$\frac{\partial}{\partial x_j} \frac{e^{x_i}}{\sum_k e^{x_k}} = s_i (1 - s_j)$$

$$\frac{\partial}{\partial x_j} s_i = \frac{\partial}{\partial x_j} \frac{e^{x_i}}{\sum_k e^{x_k}} = \frac{-e^{x_i} \cdot e^{x_j}}{(\sum_k e^{x_k})^2} = -s_i \cdot s_j$$

כעת, נרמז $j \neq i$

$$J_x(s)_{i,j} = \begin{cases} s_i (1 - s_j) & i=j \\ -s_i \cdot s_j & i \neq j \end{cases}$$

: (ב) נציגנו פורם

2.2 linear regression

: 1 מינימום

: רצונה הינה ב כווניות (a)

הוכחה \Leftarrow : יהי $v \in \ker(X)$

$$\ker(X^T X) \ni v \text{ כיון } X^T X v = X^T(Xv) = X^T 0 = 0$$

הוכחה \Rightarrow : יהי $v \in \ker(X^T X)$

: $Xv = w$: $w \neq 0$ כי $v \in \ker(X) \Rightarrow v \in \ker(X^T)$ עליה

$$X^T X v = 0$$

$$\Rightarrow v^T X^T X v = v^T 0 = 0$$

$$\Rightarrow (Xv)^T X v = 0$$

$$\Rightarrow w^T w = 0$$

$$\Rightarrow \|w\|^2 = 0$$

כך v מושך מינימום רצוניה כפניהם

$v \in \ker(X)$, \perp

: רצונה הינה ב כווניות (b)

הוכחה \Rightarrow : $A^T w = v$ כי $v \in \ker(A)$

: $0 = \langle u \cdot v \rangle = \langle A u \cdot v \rangle$ כי $A u = 0$ כי $v \in \ker(A)$

$$\langle v \cdot u \rangle = \langle A^T w \cdot u \rangle = \langle w \cdot A u \rangle = \langle w \cdot 0 \rangle = 0$$

. $\ker(A)^\perp \ni v \Rightarrow v \in V$ כי $v \in \ker(A)^\perp$

הוכחה \Leftarrow : $\mathbb{R}^n = \ker(A) \oplus \ker(A)^\perp$ כי $\ker(A) \cap \ker(A)^\perp = \{0\}$

. $w_2 \in \ker(A)^\perp$ כי $\ker(A)^\perp \perp \ker(A)$ כי $w = w_1 + w_2$ כי $w \in \ker(A)^\perp$

לפיכך $w \in \ker(A)^\perp$ כי $w = 0 + w$ כי $w \in \ker(A)^\perp$

. $\text{Im}(A^T) \supseteq \ker(A)^\perp$ כי $\ker(A)^\perp \subseteq \text{Im}(A^T)$ כי $\ker(A)^\perp \subseteq \text{Im}(A^T)$

$y \in \ker(x^T) \Rightarrow y \in \text{Im}(x)^\perp$ כי $y \perp \ker(x^T)$ כי $\ker(x^T) \subseteq \text{Im}(x)^\perp$

כי $\ker(A)^\perp = \text{Im}(A^T)^\perp$ כי $\ker(A)^\perp \subseteq \text{Im}(x)^\perp$ כי $\ker(A)^\perp \subseteq \text{Im}(x)$

לפיכך $\ker(A)^\perp = \text{Im}(x)$ כי $\dim(\ker(A)) = \dim(\text{Im}(x))$

ויהי $y = Xw$ מילולית כי $\mathbb{R}^n \neq \text{Im}(x)$ כי $\ker(x^T) \neq \text{Im}(x)$

פיזיולוגיים.

הוכחה \Leftarrow : $y = Xw$ מילולית כי $y \in \text{Im}(x)$

. $\ker(x) \ni u \Rightarrow x^T u = 0$ כי $x^T u \in \ker(x^T)$

$x^T u = 0$ כי $x^T u = 0$ כי $x^T u = 0$

שי $y = Xw_0$ כי $y \in \text{Im}(x)$

$$\langle Xw_0, u \rangle = \langle w_0, x^T u \rangle = \langle w_0, 0 \rangle = 0$$

. $\ker(x^T) \ni u$ כי $y \in \text{Im}(x)$

(ב) רצוי $x^T x$ הוכח כי $(x^T x)^{-1}$ מקיים:

$$x^T x w = x^T y \Rightarrow (x^T x)^{-1} (x^T x) w = (x^T x)^{-1} x^T y$$

$$\Rightarrow w = (x^T x)^{-1} x^T y$$

לזה הפטון היחי, NCII, $\mathbf{x}^T \mathbf{x}$ הפכה וכך הגדלה אלה
שווים ג-ה, גורר, כי הפטון היגייני. וכך נקבעו
זיהוי, ואנו הפטון היחי.

0 ≠ ker($x^T x$) הוכיחו: כי $x^T x$ סדרתי, כמו כן $x^T x \in \mathbb{R}$, כלומר $x^T x = 0$! $x^T x u = 0 -\ell$ כי $\ell \in \ker(x^T x)$ ו- $w_0 + u = w$, כי $x^T x w = x^T y$ כי $x^T x w \in \ker(x^T x)$, $w = w_0 + u$ כי $x^T x w = x^T x (w_0 + u) = x^T x w_0 + x^T x u = x^T y$: ב- (1) פהו נקבע כי $x^T x u = 0$ כי $x^T x u = 0$

2.2.2 Least Squares

: P"גונ X = UΣV^T . x le SVD פולינומית ליניארית

$$\begin{aligned}
 [x^T x]^{-1} x^T &= [(v \Sigma v^T)^T (v \Sigma v^T)]^{-1} (v \Sigma v^T)^T \\
 &= [v \Sigma^T v^T v \Sigma v^T]^{-1} (v \Sigma^T v^T) \\
 &= [v \Sigma^T \Sigma v^T]^{-1} (v \Sigma^T v^T) \\
 &= v (\Sigma^T \Sigma)^{-1} v^T v \Sigma^T v^T \\
 &= v D^{-1} \Sigma^T v^T
 \end{aligned}$$

$$D_{ii} = \sigma_i^2 : \text{ר' ראנליזיס מילינרי וטכני (ב) } D = \Sigma^T \Sigma \text{ כוכב}$$

: ר' ראנליזיס, פוליפ

$$[D^{-1}\Sigma^T]_{ii} = \frac{1}{\sigma_i^2} \sigma_i = \frac{1}{\sigma_i} = \Sigma_{ii}^+$$

הוכחה: ימינה שORTHOGONALITY גוררת:

$$[X^T X]^{-1} X^T y = V \Sigma^+ U^T y = X^+ y$$

Practical part – IML

3.1 Fitting A Linear Regression Model

1. Which features to keep and which not?

בשלבי העיבוד המקדים, בחרתי להוריד את העמודות הבאות: "Id", "date", "lat" "long" .

מקחת אוטם כי "תעודת זהות" של נכס ו"תאריך" הם מזהים שאינם רלוונטיים לחיזי המחר מאחר שהם אינם משקפים דבר על הנכס עצמו אלא הם נתונים סידוריים פנימיים של בסיס הנתונים. ה"lat"-ו"long"- הם מידע גיאוגרפי שאינם תורם לכוח הנבי של המודל. אומנם, מיקום גיאוגרפי של נכס מהוות מרכיב משמעותי בקביעת המחר, אך לא ברמת המיקום על פני כדור הארץ, אלא יותר נקודתי – ברמת האזור (תוכנות המיקוד).

2. Which features are categorical and how did you treat them?

המאפיינים הקטגוריאליים הינם "grade" ו "waterfront", "view", "condition" רשומות בהן היו ערכים שאינם נמצאים בטוחים של כל אחת מן התכונות, הסרתי את השורות האלה.

3. What other features did you design and what is the logic behind creating them?

הוספה عمودה בשם "years_since_renovation" . תכמה זו נועדה להיעיד על מספר השנים מאז שנכס עבר שיפוץ כלשהו. (במידה והבית לא עבר שיפוץ, הערך 0 ניתן לתוכנה זו המעד שהבית לא עבר שיפוץ).

4. How did you treat invalid/missing values?

עבור ערכים לא חוקיים או חסרים פעלתי באופן הבא:

הסרתי שורות כפולות. שורות בהן היו חסרים חלק מן הערכים, הסרתי אותן מטבלת הנתונים.

shorest בנה שנת שיפוץ הבית הייתה קטן יותר משנה הבנייה של הביתoso גם הוא. רשומות בהן היו ערכים שאינם הגיוניים הוסרו. לדוגמה, עבור `bedrooms/bathrooms/sqft_above/sqft_basement/yr_renovated` בדקתי שהערכים הם לפחות גודלים ממש מאפס.

בנוסף, עבור התכונות הקטגוריאליות, בדקתי שהערכים בטוחים המתאימים. מידע שבדקתי באינטרנט, מצאתי שעבור `waterfront` הערכים התקיימים הם {0,1}. עבור `view` הערכים התקיימים הם [0,4]. עבור `condition`, הערכים התקיימים הם [1,5] ועבור `grade`, הערכים התקיימים הם [1,13]. ע"כ שורות שהערכים לא היו בטוחים התקיימים, הוסרו.

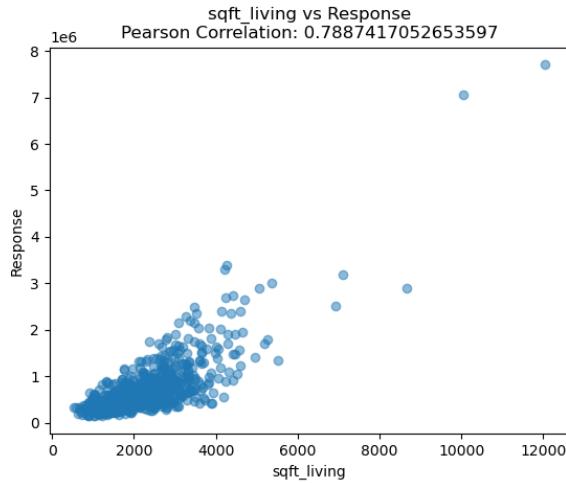
5. Explain any additional processing performed on the data.

טיפול בשורות ללא שיפוץ: עבור התכמה החדש `yrs_since_renovation`, שורות שהן `yr_renovated = 0` (המציאן שאין שיפוץ) מוגדרות באופן ספציפי ל-0 עבור `yrs_since_renovation`.

Pearson Correlation

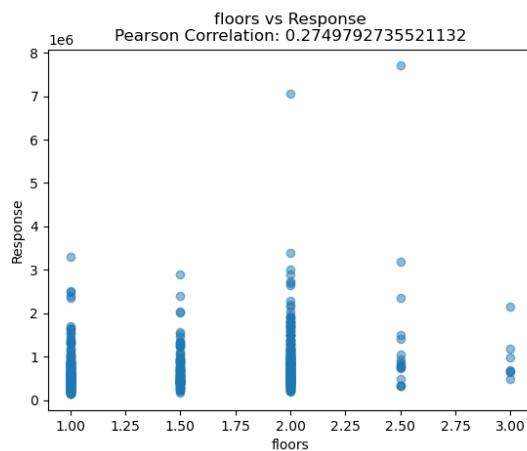
הפייצרים אותם בחרתי הם: floors | sqft_living

נסתכל על sqft_living



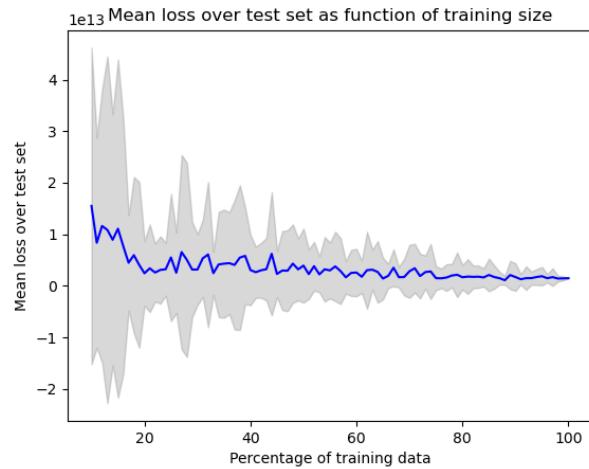
נראה כי מתאם הפירסון הוא גבוה וקרוב מאוד לאחד כלומר, ישנה קורלציה גבוהה וلينארית בין הגודל השטח של הנכס לבין מחירו. ככל שערך התכונה גבוה, מחיר הדירה עולה. מכך, ניתן להסיק שככל שטח הדירה גדול יותר כך המחיר של הדירה עולה בהתאם.

נסתכל על floors:



נראה כי מתאם הפירסון הוא נמוך יחסית עד כדי אף, כולל זיהוי תכונה שאינה תורמת למודל. לא קיים קשר לינארי כלשהו בין מספר הקומות לדירה לבין המחיר הדירה. על כן, המודל אינו יכול להסביר וללמוד מתכונה זאת ולתת חיזוי לכלהו למחיר הדירה.

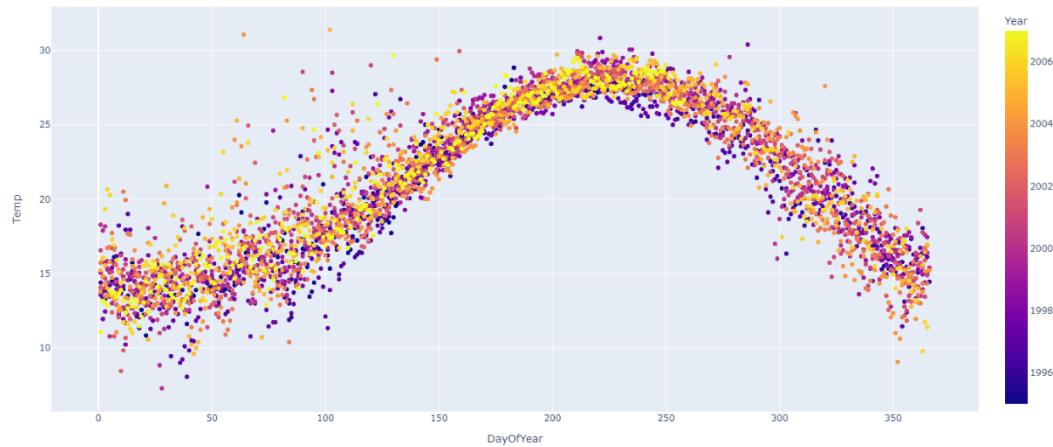
Fit a linear regression model over increasing percentages .6



ניתן לראות שעקבות MSE עבור ערכים קטנים, היא בוגמת ירידה אבל הצל משלב מסוים היא מתיצבת. בנוספ', נשים לב שככל שsett האימון הולך וגדל ה MSE הולך וקטן, במילאים אחרות, הטעות בין החיזוי של המודל למחר ולבין המחיר האמיתי הולך ונעשה קטן ככלsett האימון גדול יותר.

3.2 Polynomial Fitting

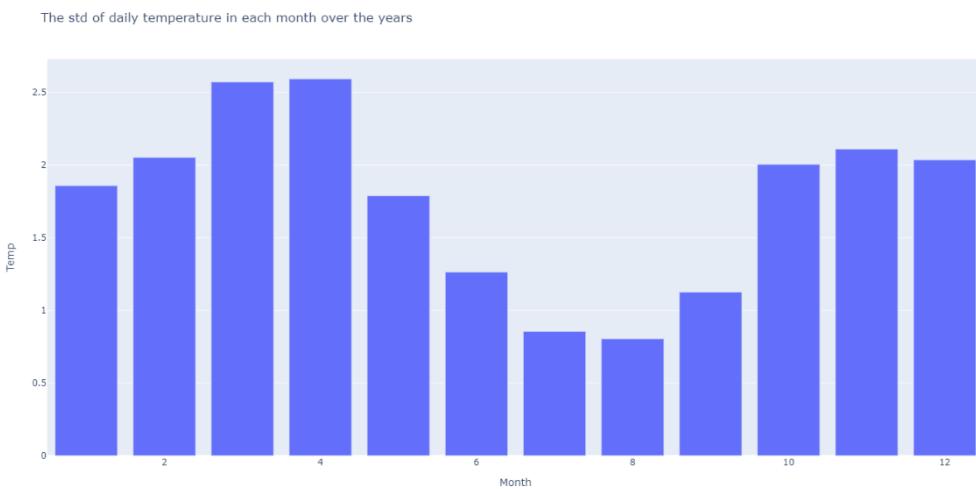
the relation between the day of the year and the temperature in Israel



3. What polynomial degree might be suitable for this data?

כפי שניתן לבדוק הימים הראשונים של השנה הטמפרטורות נמוכות וככל שמתקרב הקיץ, הטמפרטורות עלות ואז שוב יונה ירידה עם סיום השנה. נשים לב שכקודת הקיצון היא בטוחה שבין 200-250 יום, אז הטמפרטורה נעה בין 25-30. מהסקה מהגרף, נצפה שפולינום ממולה 3 לפחות יספיק.

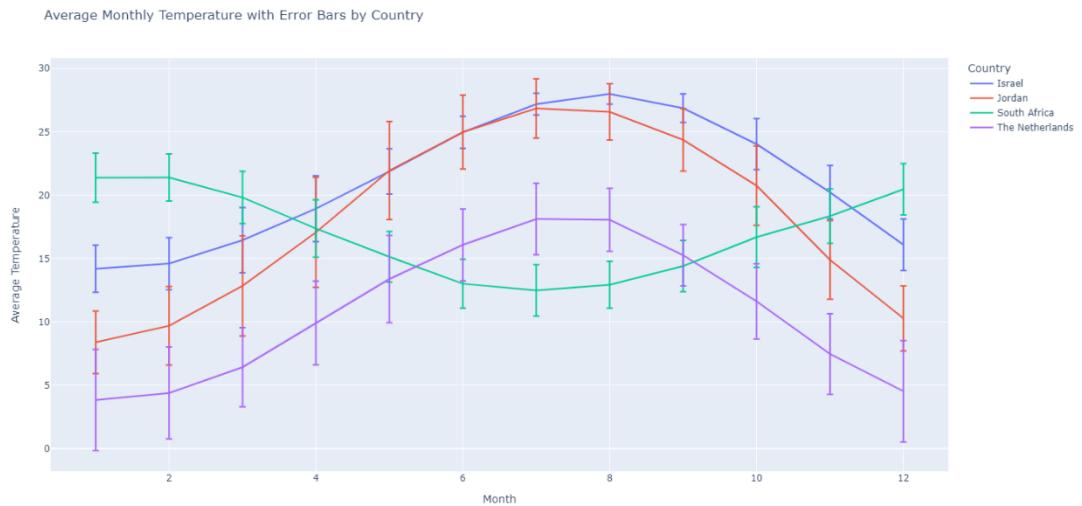
Based on this graph, do you expect a model to succeed equally over all months or are there times of the year where it will perform better than on others?



נבחן בגרף ונשים לב שעבור חודשים 9-6 סטיית התקן נמוכה מאשר יתר החודשים בשנה.

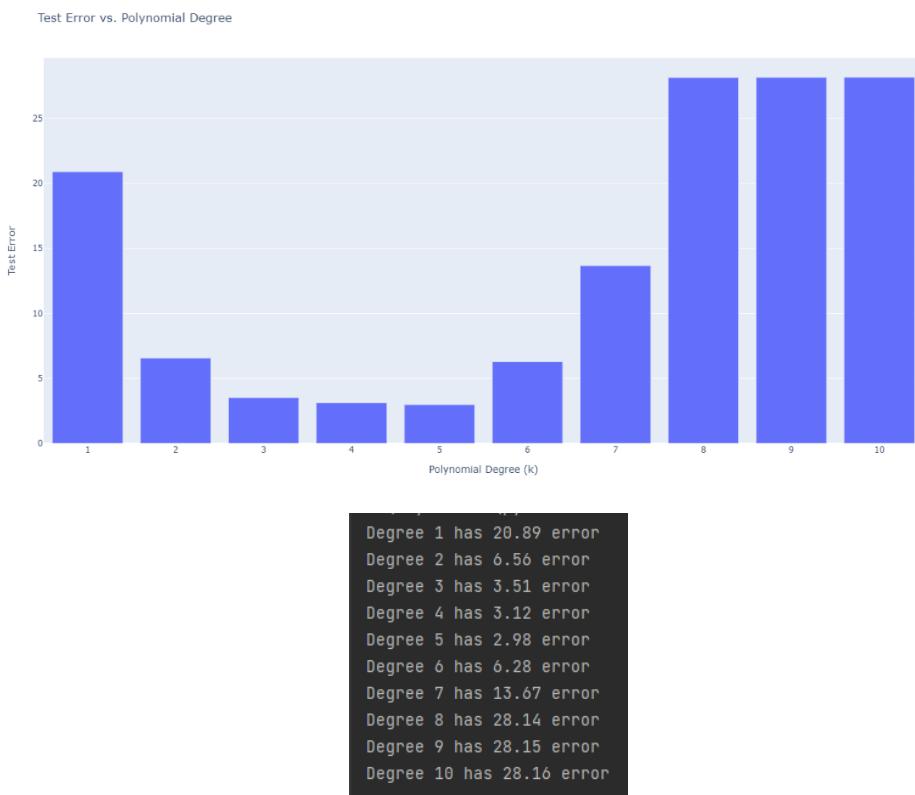
המודל יוכל לחזות בהצלחה יותר עבור חודשים 9-6 בסטיית התקן היא נמוכה, לעומת חודשים 1-8. ובאופן דומה, עבור חודשים שבהם סטיית התקן היא גבוהה, נצפה שהחיזוי יהיה פחות מוצלח.

4. Based on this graph, do all countries share a similar pattern? For which other countries is the model fitted for Israel likely to work well and for which not? Explain your answers



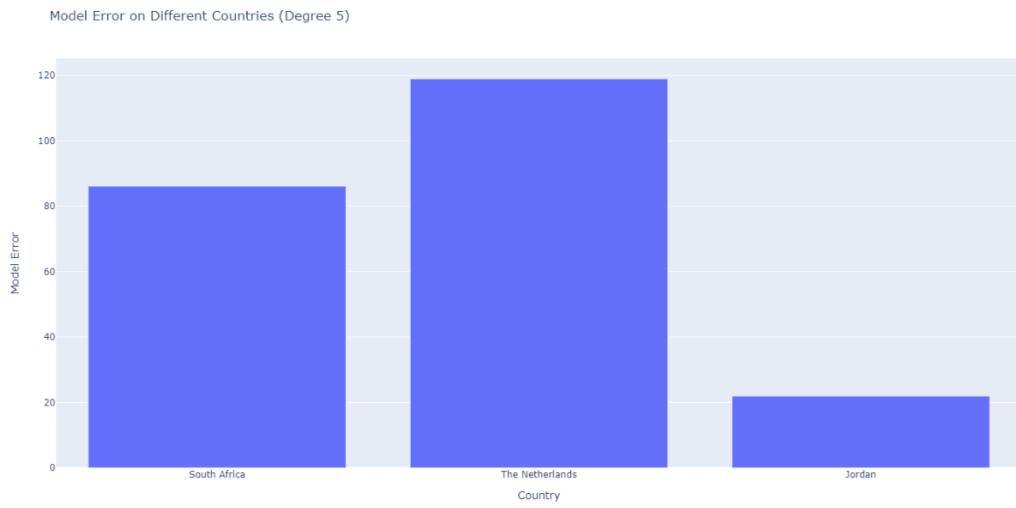
המגרף ניתן לראות שלא כל המדינות חולקות את אותה התבנית. נשים לב שמדינת ירדן כן חולקת עם ישראל תבנית דומה ועל כן נזכה שהמודל יחזז בהצלחה את הטמפרטורות שבה. בנוסף, הולנד דומה לישראל אך מוסטת בכ+ - 10 מעלות. דרום אפריקה שונה מהtbנית של ישראל ואני חולקתtbנית כלשהי אליה על כן נזכה שעבורה המודל לא יתpn פרדייקציות טובות כלשהן.

5. 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?



המגרף ומההדפסות ניתן להבחן כי הדרגה המיטבית ביותר של הפולינום בה המודל נותן את הפרדיקציה עם הטעות המינימלית היא כאשר $k=5$.

6. Plot a bar plot showing the model's error over each of the other countries



כפי ניתן להסיק מהגרף, טעות החיזוי הנמוכה ביותר התקבלה עבורה ירדן כמצופה. וזאת ממש שакן ירדן בעלת תבנית דומה מאוד לישראל ועל כן המודל הצליח לחזות פרדיקציות טובות מאוד עבורה.

נשים לב שאף על פי שהולנד חולקת התבנית דומה לישראל אך הטמפרטורות בה מושטות בכ-10 מעלות מישראל, המודל לא הצליח לחזות עבורה במדויק מיטבית ואף גודל הטעות עבורה הינו הגadol ביותר מ בין שלושת המדינות.

ניתן לשער שהחיזוי עבור דרום אפריקה שטוב יותר מאשר מושטוטים של הולנד על אף השוני המוחלט בין התבניתה לבין התבנית של ישראל נובע מכך שהטמפרטורות בדרום אפריקה דומות לטמפרטורות של ישראל ועל כן המודל מוציא תוצאות טובות יותר עבור דרום אפריקה מאשר להולנד, אך עדין הטעות גבוהה ממש מוגעתית מירדן, המהווה את ההתאמה הכי קרובה לישראל.