

2.1 Solutions of The Normal Equations

Based on Lecture 2 and Recitation 3

- Prove that: $\text{Ker}(\mathbf{X}) = \text{Ker}(\mathbf{X}^\top \mathbf{X})$
- Prove that for a square matrix A : $\text{Im}(A^\top) = \text{Ker}(A)^\perp$
- 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 $\Leftrightarrow \mathbf{y} \perp \text{Ker}(\mathbf{X}^\top)$.
- 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).

$$\text{ker}(\mathbf{x}) = \text{ker}(\mathbf{x}^\top \mathbf{x})^\perp \quad (1)$$

$$\forall v \in \text{ker}(\mathbf{x}^\top \mathbf{x}) \iff (\mathbf{x}^\top \mathbf{x})v = \mathbf{x}^\top (\mathbf{x}v) = \mathbf{x}^\top \bar{0} = \bar{0} \quad \text{because } \mathbf{x}v = \bar{0} \quad \forall v \in \text{ker}(\mathbf{x}) \iff \mathbf{x}^\top \mathbf{x}v = 0 \quad \text{because } \mathbf{x}^\top \mathbf{x}v = 0 \Rightarrow v \in \text{ker}(\mathbf{x}^\top \mathbf{x})$$

$$\mathbf{x}^\top \mathbf{x}v = \bar{0} \Rightarrow v^\top \mathbf{x}^\top \mathbf{x}v = 0 \Rightarrow (\mathbf{x}v)^\top \mathbf{x}v = 0 \iff \langle \mathbf{x}v, \mathbf{x}v \rangle = 0 \Rightarrow \mathbf{x}v = \bar{0} \Rightarrow v \in \text{ker}(\mathbf{x})$$

$$\therefore \text{Im}(A^\top) = \text{ker}(A)^\perp - \{0\}, \text{ and } A = \begin{pmatrix} -a_1 & \dots \\ \vdots & \ddots \\ -a_n & \end{pmatrix} \in \mathbb{R}^{m \times n}, \text{ no } (2)$$

$$\text{Im}(A^\top) = \text{span}(a_1, \dots, a_n)^\perp \iff \forall 1 \leq i \leq n \quad \langle a_i, x \rangle = 0 \iff A \cdot x = \begin{pmatrix} -a_1 & \dots \\ \vdots & \ddots \\ -a_n & \end{pmatrix} \cdot x = \bar{0} \iff x \in \text{ker}(A)$$

$$\text{ker}(A) \subseteq \text{Im}(A^\top) \quad \text{because } x \in \text{Im}(A^\top)^\perp \iff x \in \text{ker}(A)^\perp \iff \langle a_i, x \rangle = 0 \quad \forall i \in [n] \quad (A u) = \langle a_i, u \rangle \quad \forall u \in \text{Im}(A^\top)^\perp \iff \langle a_i, u \rangle = 0 \quad \forall u \in \text{ker}(A) \quad \text{because } A u = 0 \iff \langle a_i, u \rangle = 0$$

$$\text{therefore } \text{Im}(A^\top) = \text{ker}(A)^\perp \text{ and } \text{Im}(A^\top)^\perp = \text{ker}(A) \text{ st}$$

- 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 $\Leftrightarrow \mathbf{y} \perp \text{Ker}(\mathbf{X}^\top)$.

(3)

$$(A^\top = x \text{ because } \text{Im}(x) = \text{ker}(x^\top)^\perp \text{ and } \text{Im}(x) \text{ is the null space of } x^\top)$$

$$\text{and } x^\top \mathbf{y} = \mathbf{y}^\top x \text{ because } \text{Im}(x) = \text{ker}(x^\top)^\perp \text{ and } \text{Im}(x) \text{ is the null space of } x^\top$$

$$\mathbf{y} \perp \text{Ker}(x^\top) \iff y \in \text{Im}(x) = \text{ker}(x^\top)^\perp \iff x^\top y = 0 \iff \mathbf{y}^\top x = 0 \iff x^\top \mathbf{y} = 0 \iff \mathbf{y} \in \text{ker}(x^\top)^\perp \text{ and } x^\top \mathbf{y} = 0$$

- 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).

(4)

$$\mathbf{w} = (x^\top x)^{-1} x^\top \mathbf{y} = x^{-1} (x^\top)^{-1} x^\top \mathbf{y} = x^{-1} \mathbf{y} - \text{proj}_{\text{Im}(x)} \mathbf{y} \quad \text{because } x^\top x \text{ is invertible}$$

because $x^\top x$ is invertible (3.10) $\Leftrightarrow x^\top x$ is full rank, $\text{Im}(x) = \text{ker}(x^\top)^\perp$, $x^\top x$ is invertible

$$x^\top \mathbf{y} \in \text{ker}(x)^\perp \text{ because } \text{Im}(x) = \text{ker}(x^\top)^\perp \text{ and } x^\top \mathbf{y} \in \text{Im}(x) \text{ because } x^\top \mathbf{y} \in \text{Im}(x^\top)^\perp \text{ and } x^\top \mathbf{y} \in \text{Im}(x)$$

because $x^\top x$ is invertible (3.10) $\Leftrightarrow x^\top x$ is full rank, $\text{Im}(x) = \text{ker}(x^\top)^\perp$, $x^\top x$ is invertible

$$\text{proj}_{\text{Im}(x)} \mathbf{y} = \text{proj}_{\text{Im}(x)} (x^\top x)^{-1} x^\top \mathbf{y} = (x^\top x)^{-1} x^\top \mathbf{y} \quad \text{because } x^\top x \text{ is invertible}$$

2.2 Projection Matrices

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).

Prove the following properties in any order you wish:

- Show that P is symmetric.
- Prove that the eigenvalues of P are 0 or 1 and that $\mathbf{v}_1, \dots, \mathbf{v}_k$ are the eigenvectors corresponding to the eigenvalue 1.
- Show that $\forall \mathbf{v} \in V P\mathbf{v} = \mathbf{v}$.
- Prove that $P^2 = P$.
- Prove that $(I - P)P = 0$.

$$\text{Proof.} \quad \text{Given } \mathbf{v}_i \mathbf{v}_i^\top \text{ is N.I.D., } (\mathbf{v}_i \mathbf{v}_i^\top)^\top = (\mathbf{v}_i^\top)^\top \mathbf{v}_i^\top = \mathbf{v}_i \mathbf{v}_i^\top \quad \text{for } 1 \leq i \leq k \quad \text{So } P \text{ is symmetric.} \quad (a)$$

$$P = \sum_{i=1}^k \mathbf{v}_i \mathbf{v}_i^\top \quad \text{is the sum of } k \text{ diagonal matrices.} \quad (b)$$

$$P \cdot \mathbf{v} = \left(\sum_{i=1}^k \mathbf{v}_i \mathbf{v}_i^\top \right) \sum_{j=1}^k \alpha_j \mathbf{v}_j = \sum_{i=1}^k \alpha_i \mathbf{v}_i \mathbf{v}_i^\top \mathbf{v}_j = \sum_{i=1}^k \alpha_i \mathbf{v}_i = \mathbf{v} \quad (c)$$

$$P \cdot \mathbf{v} = \left(\sum_{i=1}^k \mathbf{v}_i \mathbf{v}_i^\top \right) \sum_{j=1}^k \alpha_j \mathbf{v}_j = \sum_{i=1}^k \alpha_i \mathbf{v}_i \mathbf{v}_i^\top \mathbf{v}_j \downarrow$$

- Since $\mathbf{v}_i \mathbf{v}_i^\top$ is O.D.N. ($\mathbf{v}_1, \dots, \mathbf{v}_k$)

$$\mathbf{v}_i^\top \mathbf{v}_j = \langle \mathbf{v}_i, \mathbf{v}_j \rangle = 0 \quad \text{if } i \neq j \quad \text{and } 1 \text{ if } i = j$$

$$\mathbf{v}_i^\top \mathbf{v}_j = \langle \mathbf{v}_i, \mathbf{v}_j \rangle = \|\mathbf{v}_i\|^2 = 1 \quad \text{if } i = j \quad \text{and } 0 \text{ if } i \neq j$$

$$V \text{ is the range space of } A^t. \quad \text{Im}(A^t)^\perp = \text{Ker}(A) \quad \text{by R.R.C.} \quad \text{Im}(A^t) = \text{Ker}(A)^\perp \quad -L \text{ is the null space of } A^t. \quad (b)$$

$$V \oplus V^\perp = V \quad \forall v \in V$$

$$\mathbb{R}^d = \text{Im}(P) \oplus \text{Im}(P)^\perp = \text{Im}(P) \oplus \text{Im}(P^t)^\perp = \text{Im}(P) \oplus \text{Ker}(P) \quad \text{by R.R.C.} \quad P \text{ is an orthogonal projection matrix.} \quad (c)$$

$$V = \text{Im}(P) \oplus \text{Ker}(P)$$

$$\text{Since } P \cdot u = \sum_{i=1}^k \mathbf{v}_i \mathbf{v}_i^\top u = \sum_{i=1}^k \langle \mathbf{v}_i, u \rangle \mathbf{v}_i \in V \Rightarrow \text{Im}P \subseteq V \quad \left. \begin{array}{l} \\ \end{array} \right\} \Rightarrow \text{Im}P = V$$

$$\forall v \in V \quad P \cdot v = v \quad \text{R.R.C.} \quad \Rightarrow \quad V \subseteq \text{Im}P$$

$$V \subseteq V_1, \dots, V_k \text{ are the null spaces of } \text{Ker}(P) \quad \text{so } V = V \oplus \text{Ker}(P) \quad -P \text{ is an orthogonal projection matrix.} \quad (d)$$

$$PV = 0 \quad \text{R.R.C.} \quad \text{if } v \in \text{Ker}(P) \quad \text{then } v \in V \quad \text{so } (c) \text{ holds.} \quad \text{R.R.C.} \quad \text{if } v \in V \text{ then } v \in V \text{ and } v \in \text{Ker}(P) \quad \text{so } v \in \text{Ker}(P) \quad \text{and } v \in V \text{ so } v = 0$$

$$PV = 0$$

$$P^2V = P(PV) = PV \quad \text{C.R.C.} \quad PV \in \text{Im}P = V \quad \text{so } V \subseteq \text{Im}P \quad (e)$$

$$P^2e_i = P^2e_i \quad \text{so } e_1, \dots, e_d \text{ are linearly independent.} \quad \text{so } V \subseteq \text{Im}P \quad (f)$$

$$P = P^2 \quad \text{so } P^2 = P$$

$$\text{Since } P \text{ is an orthogonal projection matrix.} \quad P(I-P) = 0 \quad \leftrightarrow \quad P^2 - P = 0 \leftrightarrow \quad P = P^2 \quad (g)$$

2.3 Least Squares

empirical risk minimization

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 Σ^\dagger is an $d \times m$ diagonal matrix, such that

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

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.
 7. Show that $\mathbf{X}^\top \mathbf{X}$ is invertible if and only if $\text{span}\{\mathbf{x}_1, \dots, \mathbf{x}_m\} = \mathbb{R}^d$.
 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 $\bar{\mathbf{w}}$, $\|\hat{\mathbf{w}}\| \leq \|\bar{\mathbf{w}}\|$.

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 $\hat{w}_i = \bar{w}_i$? What is the value of \hat{w}_i for the other coordinates? If you are not sure, go back to the derivation of $\hat{\mathbf{w}}$ (see recitation 4).

$$\begin{aligned}
 & -\text{בנוסף, } x = u \in V^T \text{ נקבע ש } (x^T x)^{-1} x^T = x^+ - l \text{ ו } \ker(x^T x) = \ker(x^+ - l) \\
 & (x^T x)^{-1} x^T = ((u \in V^T)^T (u \in V^T))^{-1} x^T = ((V \in \mathbb{C}^{n \times n})^T (V \in \mathbb{C}^{n \times n}))^{-1} x^T = (V \in \mathbb{C}^{n \times n})^{-1} V^T x^T = (V^T)^{-1} (V \in \mathbb{C}^{n \times n})^{-1} V^T x^T = \\
 & \quad \downarrow \quad \downarrow \quad \downarrow \\
 & (A^T)^{-1} = B^T A^T \quad (\rho \beta)^{-1} = B^T A^T \quad \text{המבחן } A, B \text{ מתקיים: } A^T B^T = I
 \end{aligned}$$

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

$$\text{IND. } x^t = \begin{pmatrix} | & | \\ x_1 & \cdots & x_m \\ | & | \end{pmatrix} : \text{פ'pnN}$$

$\det(x^t) = \det(x) - t$ \Leftrightarrow $x \in \ker(x)$ \Leftrightarrow $\ker(x) = \text{span}(x_1, \dots, x_d) = \mathbb{R}^d$

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 $\bar{\mathbf{w}}$, $\|\hat{\mathbf{w}}\| \leq \|\bar{\mathbf{w}}\|$.

$$\|\hat{\mathbf{w}}\| \leq \|\bar{\mathbf{w}}\| - \text{epsilon}$$

אנו מוכיחים כי $\|\hat{\mathbf{w}}\| \leq \|\bar{\mathbf{w}}\|$ אם ורק אם $\|\mathbf{X}^\dagger \mathbf{y}\| \leq \|\mathbf{X}^\dagger \bar{\mathbf{w}}\|$. כלומר, מוכיחים כי $\text{rank}(\mathbf{X}) = \text{rank}(\mathbf{X}^\dagger \mathbf{y})$

$$\mathbf{X} = \mathbf{U} \Sigma \mathbf{V}^\top = \begin{bmatrix} \mathbf{U}_r & \mathbf{U}_n \\ m \times m & m \times d \end{bmatrix} \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{V}_r^\top \\ \mathbf{V}_n^\top \end{bmatrix}_{d \times d}$$

$$\mathbf{X}^\dagger \mathbf{y} = \mathbf{X}^\dagger \mathbf{X} \bar{\mathbf{w}} \iff \mathbf{V} \Sigma^T \mathbf{U}^\top \mathbf{y} = \mathbf{V} \Sigma^T \mathbf{U}^\top \mathbf{V} \Sigma \bar{\mathbf{w}} \iff \mathbf{V} \Sigma^T \mathbf{U}^\top \mathbf{y} = \mathbf{V} \Sigma^2 \mathbf{V}^\top \bar{\mathbf{w}} \iff \Sigma^T \mathbf{U}^\top \mathbf{y} = \Sigma^2 \mathbf{V}^\top \bar{\mathbf{w}}$$

$$\Sigma^2 = \Sigma \Sigma = \begin{pmatrix} \sigma_1 & & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_r \end{pmatrix} \begin{pmatrix} \sigma_1 & & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_r \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_r^2 \end{pmatrix}$$

$$\Sigma^2 \mathbf{V}^\top \bar{\mathbf{w}} = \begin{pmatrix} \sigma_1^2 & & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{V}_R^\top \\ \mathbf{V}_N^\top \end{pmatrix} \bar{\mathbf{w}} = \begin{pmatrix} -\sigma_1^2 \mathbf{V}_n^\top \\ -\sigma_r^2 \mathbf{V}_r^\top \\ 0 \\ \vdots \\ 0 \end{pmatrix} \bar{\mathbf{w}} = \begin{pmatrix} \sigma_1^2 \langle \mathbf{v}_1, \bar{\mathbf{w}} \rangle \\ \vdots \\ \sigma_r^2 \langle \mathbf{v}_r, \bar{\mathbf{w}} \rangle \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$\Sigma^T \mathbf{U}^\top \mathbf{y} = \begin{pmatrix} \sigma_1 & & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{U}_R^\top \\ \mathbf{U}_N^\top \end{pmatrix} \mathbf{y} = \begin{pmatrix} \sigma_1 \langle \mathbf{u}_1, \mathbf{y} \rangle \\ \vdots \\ \sigma_r \langle \mathbf{u}_r, \mathbf{y} \rangle \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$\hat{\mathbf{w}} = \sum_{i=1}^d b_i \mathbf{v}_i, \quad \bar{\mathbf{w}} = \sum_{i=1}^d \alpha_i \mathbf{v}_i. \quad \text{הנורמליזציה של } \mathbf{X}^\dagger \text{ היא } \mathbf{X}^\dagger = \mathbf{V} \Sigma^{-1} \mathbf{U}^\top.$$

$$\forall i \in [r] \quad \sigma_i^2 \langle \mathbf{v}_i, \bar{\mathbf{w}} \rangle = \sigma_i \langle \mathbf{u}_i, \mathbf{y} \rangle : \text{מזהה ש } \mathbf{u}_i \text{ הוא מושג של } \mathbf{v}_i \text{ ב- } \mathbf{X}^\dagger.$$

$$\sigma_i^2 \langle \mathbf{v}_i, \hat{\mathbf{w}} \rangle = \sigma_i^2 \langle \mathbf{v}_i, \bar{\mathbf{w}} \rangle \stackrel{\sigma_i \neq 0}{\iff} \langle \mathbf{v}_i, \hat{\mathbf{w}} \rangle = \langle \mathbf{v}_i, \bar{\mathbf{w}} \rangle \iff \langle \mathbf{v}_i, \sum_{i=1}^d b_i \mathbf{v}_i \rangle = \langle \mathbf{v}_i, \sum_{i=1}^d \alpha_i \mathbf{v}_i \rangle \iff \alpha_i = b_i$$

$$\text{וככלומר } \sum_{i=1}^d \alpha_i = \sum_{i=1}^d b_i.$$

$$\forall i \neq j \in [d] \quad \langle \mathbf{v}_i, \mathbf{v}_j \rangle = 0 : \text{מכיוון ש } \mathbf{v}_i \text{ ו } \mathbf{v}_j \text{ אינטראקציית לא-}$$

$$\langle \mathbf{v}_i, \mathbf{v}_i \rangle = 1$$

$$\text{בנוסף, } \mathbf{v}_i \text{ ו } \mathbf{v}_j \text{ אינטראקציית לא-}$$

$$\hat{w} = X^T y = V \Sigma^T V^T y = \begin{pmatrix} | & | & | \\ | & \dots & | \\ | & | & | \end{pmatrix} \begin{pmatrix} \frac{1}{\sigma_1} & 0 & 0 \\ 0 & \frac{1}{\sigma_r} & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} -u_1 \\ \vdots \\ -u_r \\ \vdots \\ -u_d \end{pmatrix} y =$$

$$\begin{pmatrix} | & | & | \\ | & \dots & | \\ | & | & | \end{pmatrix} \begin{pmatrix} \frac{1}{\sigma_1} & 0 & 0 \\ 0 & \frac{1}{\sigma_r} & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \langle u_1, y \rangle \\ \vdots \\ \langle u_r, y \rangle \\ \vdots \\ \langle u_d, y \rangle \end{pmatrix} = \begin{pmatrix} | & | & | \\ | & \dots & | \\ | & | & | \end{pmatrix} \begin{pmatrix} \frac{\langle u_1, y \rangle}{\sigma_1} \\ \vdots \\ \frac{\langle u_r, y \rangle}{\sigma_r} \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \sum_{i=1}^r \frac{\langle u_i, y \rangle}{\sigma_i} v_i$$

↓

$\forall i \in \{r+1, \dots, d\} \quad b_i = 0$

$$\|\hat{w}\|^2 = \langle \hat{w}, \hat{w} \rangle = \left\langle \sum_{i=1}^r b_i v_i, \sum_{i=1}^r b_i v_i \right\rangle = \sum_{i=1}^r b_i^2 \langle v_i, v_i \rangle = \sum_{i=1}^r b_i^2 = \sum_{i=1}^r a_i^2$$

: $\rho \in \mathbb{R}^{N \times N}$

$\sum_{i=1}^r b_i^2$
 $\|v_i\| = 1$
 $a_i = b_i$
 $i \in [r]$

$\sum_{i=r+1}^d b_i^2 = 0$
 $\langle v_i, v_i \rangle = 0$
 $v_i \in \mathbb{R}^N$
 $v_i \perp v_j \quad \forall i \neq j$

$$\|\bar{w}\| = \langle \bar{w}, \bar{w} \rangle = \left\langle \sum_{i=1}^d a_i v_i, \sum_{i=1}^d a_i v_i \right\rangle = \sum_{i=1}^d a_i^2$$

- מינימום נורמה \bar{w} ב \mathbb{R}^d

$\sum_{i=1}^d a_i^2$
 $\|v_i\| = 1$
 $a_i \in \mathbb{R}$

$$\text{האנו} \quad \|\hat{w}\| \leq \|\bar{w}\| \quad \Leftarrow \quad \|\hat{w}\|^2 = \sum_{i=1}^r a_i^2 \leq \sum_{i=1}^d a_i^2 = \|\bar{w}\|^2 \quad \text{כיוון}$$

$\sum_{i=1}^r a_i^2$
 $r \leq d$

3 Practical Part

3.1 Fitting A Linear Regression Model

2. Implement the `preprocess_data` function. The function receives a loaded pandas `DataFrame` object of the `observation matrix`, and a pandas `Series` object of the `response vector`, and returns them `after preprocessing`. Make sure that if you remove a sample, you remove it properly from both the observation matrix as well as from the response vector.

Explore the data (some information can be found on [Kaggle](#)) and perform any necessary preprocessing (such as but not limited to):

- What sort of values are valid for different types of features? Can house prices be negative? Can a living room size be too small?
 - Some of the features are categorical with no apparent logical order to their values (for example zip-code). Correctly address these features such that it will make sense to fit a linear regression model using them. For assistance you may refer to the following [StackOverflow question](#).
 - Are there any additional features that might be beneficial for predicting the house price and that can be derived from existing features?

*Notice: Your code should run properly also if it does not receive a response vector (y is None). That is for the case of inference, where you will receive house features without the label.

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.

כיה נחריר הCANbus נספחים:

`id` - Unique ID for each home sold
`date` - Date of the home sale
`price` - Price of each home sold
`bedrooms` - Number of bedrooms
`bathrooms` - Number of bathrooms, where .5 accounts for a room with a toilet but no shower
`sqft_living` - Square footage of the apartments interior living space
`sqft_lot` - Square footage of the land space
`floors` - Number of floors
`waterfront` - A dummy variable for whether the apartment was overlooking the waterfront or not
`view` - An index from 0 to 4 of how good the view of the property was
`condition` - An index from 1 to 5 on the condition of the apartment
`grade` - An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high quality level of construction and design
`sqft_above` - The square footage of the interior housing space that is above ground level
`sqft_basement` - The square footage of the interior housing space that is below ground level
`yr_built` - The year the house was initially built
`yr_renovated` - The year of the house's last renovation
`zipcode` - What zipcode area the house is in
`lat` - Latitude
`long` - Longitude
`sqft_living15` - The square footage of interior housing living space for the nearest 15 neighbors
`sqft_lot15` - The square footage of the land lots of the nearest 15 neighbors

- Which features to keep and which not?

על כל הלקוחות נקבעו שדות `lat`, `long`, `name`, `city`, `zip_code`, `key` ו-`password`.
בנוסף לכך, נקבעו שדות `date` (הנאריך בו הגיע הרוכסן), `order_id` (המספר
הנאריך בו הגיע הרוכסן), `order_qty` (כמות המזון) ו-`order_time` (זמן
ההצנה).
בנוסף לכך, נקבעו שדות `lat`, `long`, `name`, `city`, `zip_code`, `key` ו-`password` בלקוחות
הקיימים. על מנת לא ליצור בעיה בעת חישוב המרחק בין שני לקוחות
הנקודות נקבעו ב-`lat` ו-`long` בלבד, ו-`name`, `city`, `zip_code` ו-`password` מוחזק
בבסיס נתונים.

- Which features are categorical how did you treat them?

בנוסף למשתנים המהווים תוצאות מודולares, נציגו בדוח הנטען גם משתנים דמויי-переменные (dummy variables) שמייצגים נטיות מסוימות. מטרת הבדיקה היא לבדוק אם קיימת נטייה מסוימת ביחס למשתנים הללו. מטרת הבדיקה היא לבדוק אם קיימת נטייה מסוימת ביחס למשתנים הללו.

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

- How did you treat invalid/missing values?

50. החלטה על הדרישות
הדרישות מוגדרת כפונקציית f שמקיימת את התכונה $f(x) = f(y)$ אם ורק אם $x = y$.
הדרישות מוגדרת כפונקציית f שמקיימת את התכונה $f(x) = f(y)$ אם ורק אם $x = y$.
הדרישות מוגדרת כפונקציית f שמקיימת את התכונה $f(x) = f(y)$ אם ורק אם $x = y$.

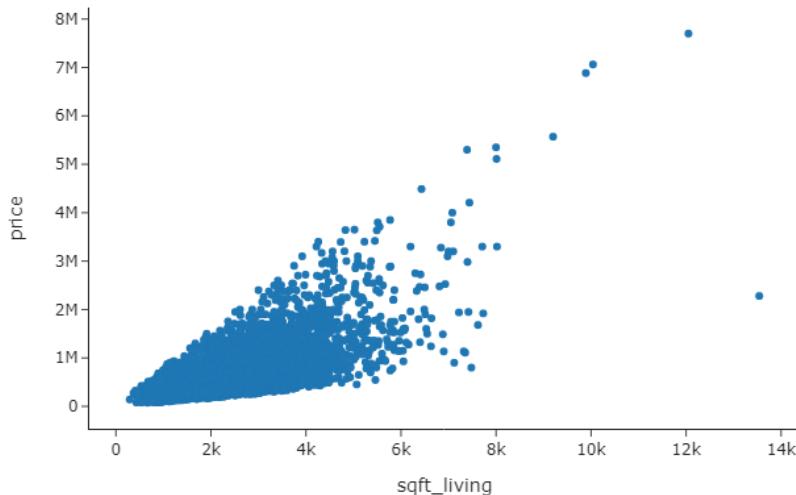
לעתה גו נתקון מ-3 גזעי נאומין פ-אט NLP. מ-3 מ-2 קי-ויאן, תי כו (null). נגיד feature-לעט לא נגיד (null), נגיד עט-ויאן לא כו (null).

- Explain any additional processing performed on the data.

הבעיר נקבעו פולחניהם הפלורר היי נציגם ותפקידם. נשים גזקיות אך גם קוויאר. אונליין או על גזק עט-ויאן אל של פ-אט' נקבעו דוחה של ה-ויאן נכון. ימיה ימי הימאות (לכ. 50% נזק), reindex-ר-עט-ויאן גזקיות דוחה נזק (לכ. 50% נזק) ו-ויאן גזקיות נזק (לכ. 50% נזק). כריסטיאן גזקיות דוחה נזק (לכ. 50% נזק) ו-ויאן גזקיות נזק (לכ. 50% נזק).

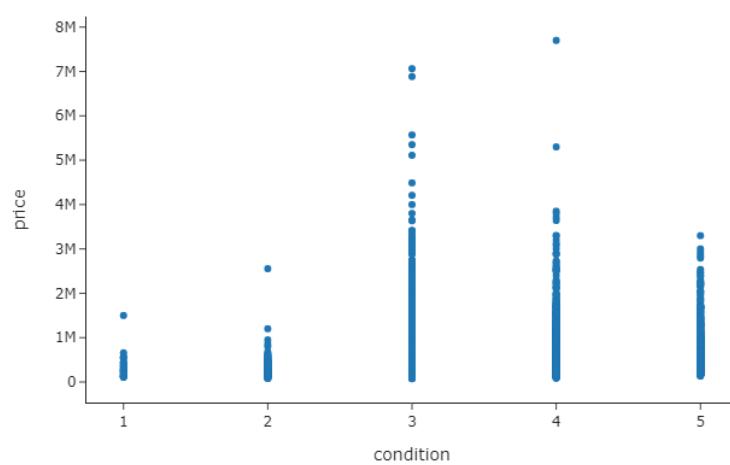
3. Choose two features, one that seems to be beneficial for the model and one that does not. In your `Answers.pdf` add the graphs of these two chosen features and explain how do you conclude if they are beneficial or not.

correlation between sqft_living and y - Pearson Correlation = 0.70241393232



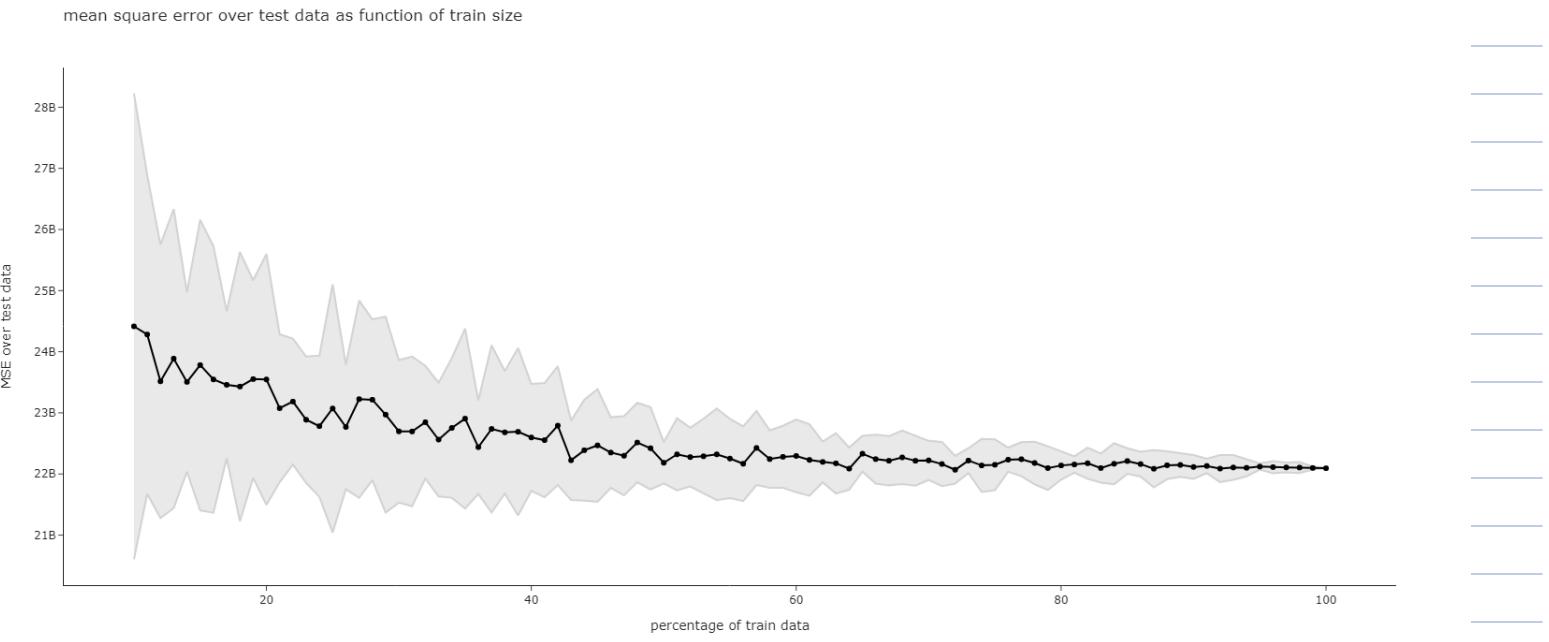
אנו נראה ש-*loss* מושג על ידי היפוך וחלוקת *soft-living* ב-*soft*.

correlation between condition and y - Pearson Correlation = 0.034277849362



הנורווגיה נבדקה נסגרת, וירען רק "ההובן" היה גורף. NCIIII אוניברסיטת נורווגיה רצתה לסייע למכון פיננסיסטי בבריטניה, נאלה וולס והסיע גראוי, אך החלטה לא נתקבלה.

4. 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?

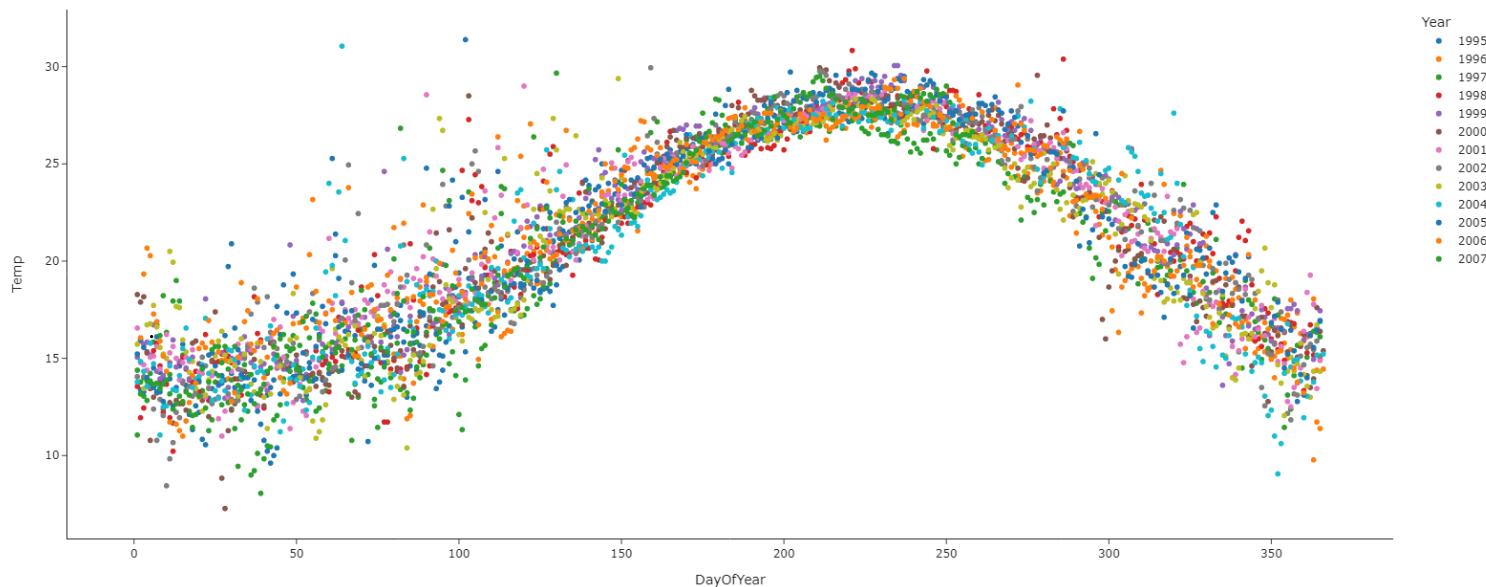


העתקים נראים בדיאגרם כטבלה של MSE וטבוקן כטבלה של גודל אינטגרל. נסמן y_i כערך מודולו קרטון. מילוי מושג y_i יתבצע באמצעות סדרת סטטיסטית $y_{i+1}, y_{i+2}, \dots, y_{i+n}$.
 רצויים שטבוקן שפונקציית ה- L_2 מינימיזציה מינימיזיר את פונקציית האנרגיה. מילוי מושג y_i יתבצע באמצעות סדרת סטטיסטית $y_{i+1}, y_{i+2}, \dots, y_{i+n}$.
 תוצאות ה- L_2 מינימיזציה מינימיזיר את פונקציית האנרגיה. מילוי מושג y_i יתבצע באמצעות סדרת סטטיסטית $y_{i+1}, y_{i+2}, \dots, y_{i+n}$.
 תוצאות ה- L_2 מינימיזציה מינימיזיר את פונקציית האנרגיה. מילוי מושג y_i יתבצע באמצעות סדרת סטטיסטית $y_{i+1}, y_{i+2}, \dots, y_{i+n}$.
 תוצאות ה- L_2 מינימיזציה מינימיזיר את פונקציית האנרגיה. מילוי מושג y_i יתבצע באמצעות סדרת סטטיסטית $y_{i+1}, y_{i+2}, \dots, y_{i+n}$.

3.2 Polynomial Fitting

What polynomial degree might be suitable for this data?

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



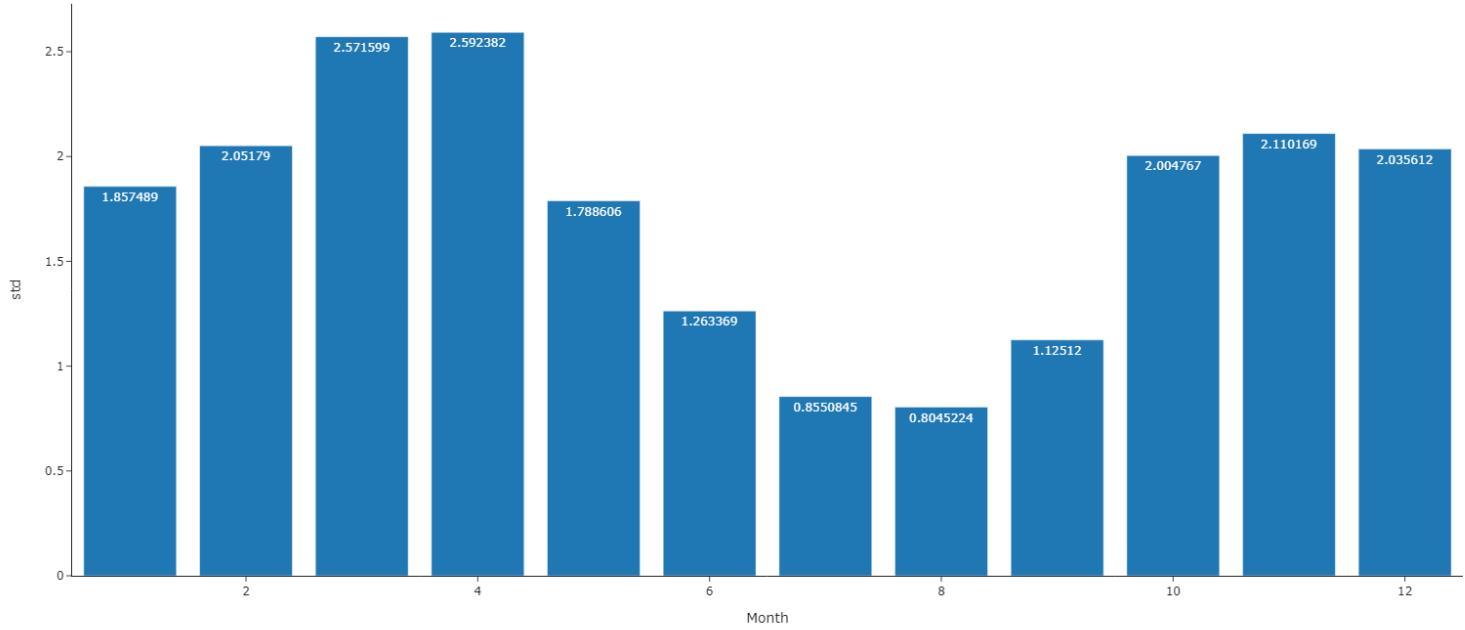
האם ניתן לאמוד מודל פולינומי עבור נתונים אלו, הינו יתאפשר וסביר? נסמן את תבנית ה- $y = f(x)$ ורוצח נאנו בוגר יותר או לא? הינה שכנן בפער נרחב בין אמצעי חישוב ומודלים (בנוסף, בפער בין תוצאות מודלים שונים).

רעיון גוף אחד נניחו ש- $f(x) = ax^3 + bx^2 + cx + d$. מילויים ש- a, b, c, d הם קבועים. אם נשים את $f(x)$ ב- $x = 0$, נקבל $f(0) = d$. אם נשים את $x = 1$, נקבל $f(1) = a + b + c + d$. אם נשים את $x = 2$, נקבל $f(2) = 8a + 4b + 2c + d$. אם נשים את $x = 3$, נקבל $f(3) = 27a + 9b + 3c + d$. אם נשים את $x = 4$, נקבל $f(4) = 64a + 16b + 4c + d$. אם נשים את $x = 5$, נקבל $f(5) = 125a + 25b + 5c + d$. אם נשים את $x = 6$, נקבל $f(6) = 216a + 36b + 6c + d$. אם נשים את $x = 7$, נקבל $f(7) = 343a + 49b + 7c + d$. אם נשים את $x = 8$, נקבל $f(8) = 512a + 64b + 8c + d$. אם נשים את $x = 9$, נקבל $f(9) = 729a + 81b + 9c + d$. אם נשים את $x = 10$, נקבל $f(10) = 1000a + 100b + 10c + d$.

הנחנו ש- $d = 0$. נזכיר ש- a, b, c, d הם קבועים. נשים את $x = 0$ ו- $f(0) = 0$, נשים את $x = 1$ ו- $f(1) = 1$, נשים את $x = 2$ ו- $f(2) = 8$, נשים את $x = 3$ ו- $f(3) = 27$, נשים את $x = 4$ ו- $f(4) = 64$, נשים את $x = 5$ ו- $f(5) = 125$, נשים את $x = 6$ ו- $f(6) = 216$, נשים את $x = 7$ ו- $f(7) = 343$, נשים את $x = 8$ ו- $f(8) = 512$, נשים את $x = 9$ ו- $f(9) = 729$, נשים את $x = 10$ ו- $f(10) = 1000$.

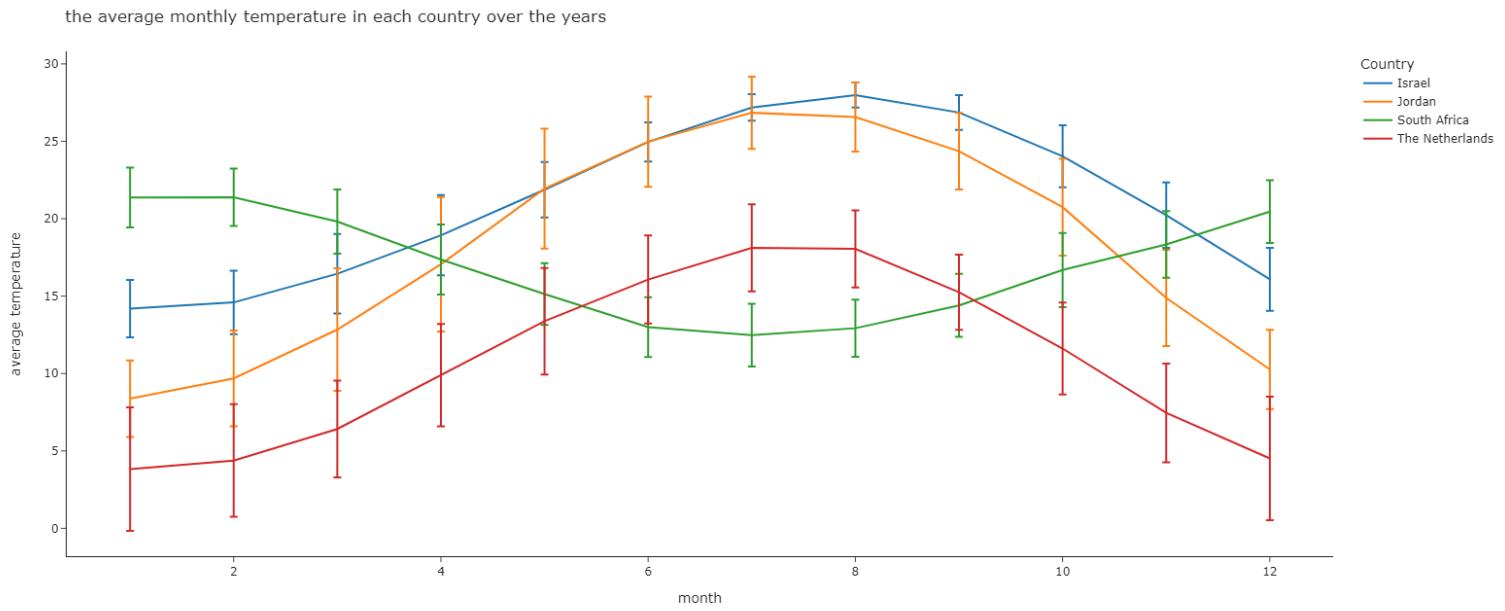
do you expect a model to succeed equally over all months or are there times of the year where it will perform better than others?

The std of daily temperature in each month over the years



ההשנה מוקדמת יותר וטמפרטורה גבוהה יותר. סטודנטים נמצאים בימי נסיעה (אפריל-מאי) וטמפרטורה גבוהה יותר. סטודנטים נמצאים בימי נסיעה (אפריל-מאי).
במשך חורף וסתיו מוקדמת יותר. סטודנטים נמצאים בימי נסיעה (אפריל-מאי).
ב-יולי מוקדמת יותר. סטודנטים נמצאים בימי נסיעות (אפריל-מאי).
בכל רגע ה-ELNINE מוקדמת יותר. סטודנטים נמצאים בימי נסיעות (אפריל-מאי).
ב-יולי מוקדמת יותר. סטודנטים נמצאים בימי נסיעות (אפריל-מאי).

3. 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.



רואים גורם אחד אחד במדינות אחרות (לעומת ישראל) פועל מינימלי (לעומת רשות).
בישראל, מינימום חורף נמוך מ-10°C, אולם גאותה של ישראל מינימום חורף נמוך מ-10°C.
בנוסף לכך, מינימום קיץ נמוך מ-20°C, אולם גאותה של ישראל מינימום קיץ נמוך מ-20°C.
בנוסף לכך, מינימום חורף נמוך מ-10°C, אולם גאותה של ישראל מינימום חורף נמוך מ-10°C.

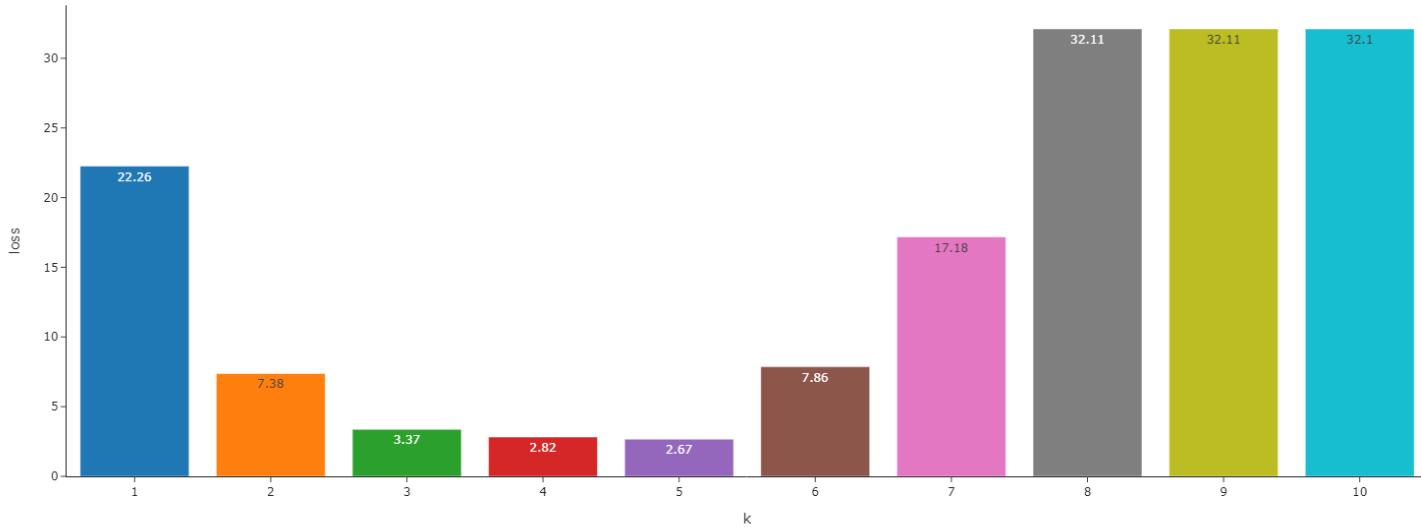
בנוסף לכך, מינימום קיץ נמוך מ-20°C, אולם גאותה של ישראל מינימום קיץ נמוך מ-20°C.

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?

the test error for degrees in range 1-10



השאלה מבקשת למצוא מודל פולינומי שמייצג את הנתונים בצורה הטובה ביותר. מטרת התרגיל היא למשוך מודלים פולינומיים מדרגות 1 עד 10 ולבזבז בפער רב במודלים הגבוהים (k=8, 9, 10) בעודם מושגים על ידי מודלים נמוכים (k=3, 4). מטרת התרגיל היא למשוך מודלים פולינומיים מדרגות 1 עד 10 ולבזבז בפער רב בפער בין מודלים נמוכים (k=3, 4) לבין מודלים גבוהים (k=8, 9, 10).

```

loss of a model with k = 1 over the test set is 22.26
loss of a model with k = 2 over the test set is 7.38
loss of a model with k = 3 over the test set is 3.37
loss of a model with k = 4 over the test set is 2.82
loss of a model with k = 5 over the test set is 2.67
loss of a model with k = 6 over the test set is 7.86
loss of a model with k = 7 over the test set is 17.18
loss of a model with k = 8 over the test set is 32.11
loss of a model with k = 9 over the test set is 32.11
loss of a model with k = 10 over the test set is 32.1

```

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.

