Exercise 2

Question 1

Suppose $A, B \in \mathbb{R}^{n\times n}$, $i,j \in \{1,\dots,n$

(a)(1) $(AB)_{ij} = \sum_{k=1}^{n} A_{ik}B_{kj} = \sum_{k=1}^{n} B^T_{jk}A^T_{ki} = (B^TA^T)_{ji}$. Hence $(AB)^T = B^TA^T$.

(2) $(A+B)^T_{ij} = (A+B)_{ji} = A_{ji} + B_{ji} = A^T_{ij} + B^T_{ij}$. Hence $(A+B)^T = A^T + B^T$.

(3) Choose $A := \begin{bmatrix} 1 & 2 \\ 0 & 3 \end{bmatrix}$, $B = \begin{bmatrix} 2 & 3 \\ 1 & 2 \end{bmatrix}$, then $AB = \begin{bmatrix} 4 & 7 \\ 3 & 6 \end{bmatrix}$ with $10 = tr(AB) \neq tr(A)\,tr(B) = 4\cdot 4 = 16$.

(4) $tr(AB) = \sum_{i=1}^{n}\sum_{k=1}^{n} A_{ik}B_{ki} = \sum_{k=1}^{n}\sum_{i=1}^{n} B_{ki}A_{ik} = tr(BA)$

(5) $tr(A+B) = \sum_{i=1}^{n}(A+B)_{ii} = \sum_{i=1}^{n}(A_{ii}+B_{ii}) = \sum_{i=1}^{n} A_{ii} + \sum_{i=1}^{n} B_{ii} = tr A + tr B$

(6) $(AB)B^{-1}A^{-1} = I$ and $B^{-1}A^{-1}(AB) = I$. Hence $(AB)^{-1} = B^{-1}A^{-1}$

(7) Choose $A := \begin{bmatrix} 1 & 2 \\ 0 & 3 \end{bmatrix}$, $B = \begin{bmatrix} 2 & 3 \\ 1 & 2 \end{bmatrix}$, then $A^{-1} = \frac{1}{3}\begin{bmatrix} 3 & -2 \\ 0 & 1 \end{bmatrix}$, $B^{-1} = \begin{bmatrix} 2 & -3 \\ -1 & 2 \end{bmatrix}$, while $A+B = \begin{bmatrix} 3 & 5 \\ 1 & 5 \end{bmatrix}$, $(A+B)^{-1} = \frac{1}{10}\begin{bmatrix} 5 & -5 \\ -1 & 3 \end{bmatrix} \neq A^{-1} + B^{-1}$

(b) Let $u, v \in \ker A$, $c \in \mathbb{R}^n$, then $A(u+v) = Au + Av = 0 + 0 = 0$, $A(cv) = cAv = 0$, and $0 \in \ker A$. Hence, $\ker A$ is a subspace of $\mathbb{R}^n$. Let $u, v \in Im A$ then $\exists u', v' \in \mathbb{R}^n$ s.t. $Au' = u$, $Av' = v$. $A(u'+v') = Au' + Av' = u+v \in Im A$ (closure under addition) and $A(cv') = cAv' = cv \in Im A$ (closure under scalar multiplication). Since $A0 = 0 \in Im A$ we get that $Im A$ is a subspace of $\mathbb{R}^n$.

(c) Suppose $A \in \mathbb{R}^{n\times p}$ with $T \in L(\mathbb{R}^p, \mathbb{R}^n)$ its linear map. Let $v \in \mathbb{R}^n$ and $u \in null\, T$, then $\langle T^*v, u\rangle = \langle v, Tu\rangle = \langle v, 0\rangle = 0$. This shows that $Im(T^*) \subseteq (null\ T)^\perp$. But $\dim (null\ T)^\perp = rank\, T = \dim Im(T^*)$, hence $Im(T^*) = (null\ T)^\perp$.

Question 2

(a) Suppose $P_U \in \mathbb{R}^{n\times n}$ is an orthogonal projection matrix, projecting onto the subspace $U \subseteq \mathbb{R}^n$. Let $u_1, \dots, u_m$ an orthonormal basis of U and extend it into an orthonormal basis $u_1, \dots, u_m, \dots, u_n$ of $\mathbb{R}^n$. $\forall v \in \mathbb{R}^n$ $\exists!\, u \in U = span(\{u_1, \dots, u_m\})$ $u' \in U^\perp = span(\{u_{m+1}, \dots, u_n\})$ s.t. $v = u + u'$. $P_U(v) = 1u + 0u'$, implying that 1 and 0 are P_U's only eigenvalues, with $\dim E(1, P_U) = m = rank\, P_U$, and $\dim E(0, P_U) = n - m = \dim null\, P_U$.

(b) Suppose $X \in \mathbb{R}^{n\times p}$ is a full rank matrix, with $T \in L(\mathbb{R}^p, \mathbb{R}^n)$ its linear map. Let $v \in \mathbb{R}^p$, then $T^*Tv = 0 \rightarrow$ $\langle T^*Tv, v\rangle = 0 \rightarrow \langle Tv, Tv\rangle = 0 \rightarrow \|Tv\| = 0 \rightarrow v = 0$ and T^*T is injective. Since it is an operator on $\mathbb{R}^p$ it is also invertible. In addition, $(T^*T)^* = T^*T$, meaning T^*T is self-adjoint. Let $P_X = T(T^*T)^{-1}T^*$, then $P_X^2 = T(T^*T)^{-1}T^*T(T^*T)^{-1}T^* = T(T^*T)^{-1}T^* = P_X$ and P_X is idempotent, meaning it is a projection. $P_X^* = (T(T^*T)^{-1}T^*)^* = T(T^*T)^{-1}T^* = P_X$, and P_X is self-adjoint. Let $v \in \mathbb{R}^n$, then $\langle P_X v, (I - P_X)v\rangle =$ $\langle P_X v, v\rangle - \langle P_X v, P_X v\rangle = \langle P_X v, v\rangle - \langle P_X^* P_X v, v\rangle = \langle P_X v, v\rangle - \langle P_X^2 v, v\rangle = 0$, and P_X is an orthogonal projection. $\forall v \in \mathbb{R}^n$ $P_X = T[(T^*T)^{-1}T^*v]$, implying that $P_X v \in Im\, T$.

(c) P_X is self-adjoint, hence according the real spectral theorem, it is diagonalizable w.r.t. some orthonormal basis

$u_1,\dots,u_n$. Let $U=[u_1,\dots,u_n]$ then $P_X = U\Lambda U^{-1}$ where Λ is an nxn diagonal matrix with p ones and $n-p$ zeros in its diagonal, corresponding to P_x's eigenvalues. $tr(P_X) = tr(U\Lambda U^{-1}) = tr(U^{-1}U\Lambda) = tr(\Lambda) = p$.

Question 3 part 1

(a) Suppose $A\in\mathbb{R}^{n\times p}$ and $T\in L(\mathbb{R}^p,\mathbb{R}^n)$ its linear map. Let $v\in\mathbb{R}^n$ s.t. $T^*Tv=\lambda v$ for some $\lambda\in\mathbb{R}$, then $TT^*Tv=\lambda Tv$, meaning that Tv is an eigenvector of TT^*, with λ being its eigenvalue. In fact, $T:E(\lambda,T^*T)\to E(\lambda,TT^*)$. For $\lambda\neq 0$ we get $Tv\neq 0$, meaning that $Tv|_{E(\lambda,T^*T)}$ is injective and $\dim E(\lambda,T^*T)\leq \dim E(\lambda,TT^*)$. Similarly, for $\lambda\neq 0$, $T^*:E(\lambda,TT^*)\to E(\lambda,T^*T)$ is injective and $\dim E(\lambda,TT^*)\leq \dim E(\lambda,T^*T)$. This implies that $E(\lambda,T^*T)=E(\lambda,T^*T)$. Let $\lambda\in\mathbb{R}$ s.t. $TT^*v=\lambda v$ for some $v\in\mathbb{R}^n$, then $\lambda\langle v,v\rangle = \langle TT^*v, v\rangle = \langle T^*v, T^*v\rangle = \|T^*v\|^2\geq 0$, implying that $\lambda\geq 0$. To conclude, T^*T and TT^* share the same non-negative eigenvalues, with the positive eigenvalues having the same geometric multiplicities.

(b) Let $r=\operatorname{rank}A=\operatorname{rank}A^T$ and assume that A can be decomposed s.t. $A=USV^T$ with $U\in\mathbb{R}^{n\times n}$ and $V\in\mathbb{R}^{p\times p}$ being orthogonal and $S\in\mathbb{R}^{n\times p}$. Then $AA^T = USV^T(USV^T)^T = USV^TVS^TU^T = USS^TU^T$, and we've got a familiar spectral decomposition of a symmetric matrix AA^T with its eigenvalues in the diagonal of $SS^T\in\mathbb{R}^{n\times n}$. Sort these eigenvalues in a descending order: $\lambda_1,\dots,\lambda_r,0,\dots,0$ and set U's columns accordingly. Similarly, $A^TA=(USV^T)^TUSV^T = VS^TSV^T$ which is the spectral decomposition of the symmetric matrix A^TA with its eigenvalues on the diagonal of $S^TS\in\mathbb{R}^{p\times p}$. According to (a), AA^T and A^TA share the same eigenvalues with positive ones having the same geometric multiplicity. Therefore $\lambda_1,\dots,\lambda_r,0,\dots,0$ are also the diagonal entries of S^TS, maybe with different number of trailing zeros, and we can sort them in a descending order and arrange V's columns accordingly. We can now define $D\in\mathbb{R}^{r\times r}$ to be a diagonal matrix with $\sqrt{\lambda_1},\dots,\sqrt{\lambda_r}$ in its diagonal, and $S\in\mathbb{R}^{n\times p}$ as $\begin{bmatrix} D & 0\\ 0 & 0\end{bmatrix}$. In order to find the singular values $s_1,\dots,s_r$ we take the square-roots of A^TA eigenvalues. V's columns $v_1,\dots,v_p$ are constructed by turning the eigenvectors into an orthonormal list, following the Gram-Schmidt procedure. Let $U=[u_1,\dots,u_n]$ and $V=[v_1,\dots,v_p]$, then $u_1,\dots,u_n$ are the orthonormal eigenvectors of TT^* and $v_1,\dots,v_p$ are the orthonormal eigenvectors of T^*T, both are relative to the same sorted eigenvalue list $\lambda_1,\dots,\lambda_r,0,\dots,0$ (up to the number of zeros). As shown in (a), $\forall i=1,\dots,r$ $T|_{E(\lambda_i,T^*T)}$ is an isomorphism between $E(\lambda_i,T^*T)$ and $E(\lambda_i,TT^*)$. But $\|Tv_i\|^2=\langle Tv_i,Tv_i\rangle = \langle v_i, T^*Tv_i\rangle = \lambda_i\|v_i\|^2=\lambda_i$ while $\|u_i\|^2=1$. Hence $u_i=\frac{Tv_i}{\sqrt{\lambda_i}}=\frac{Tv_i}{s_i}$.

part 2

(a) The first principal component, defined by the unit vector $w\in\mathbb{R}^p$ that maximizes the variance of the data points $x_1,\dots,x_n\in\mathbb{R}^p$ is formally defined as $PC_1^{var}=\underset{w\in\mathbb{R}^p,\ \|w\|=1}{\arg\max}\|Xw\|^2$ using the standard inner product space over $\mathbb{R}$ with $X=[x_1,\dots,x_n]^T$. Let $T\in(\mathbb{R}^p,\mathbb{R}^n)$ the corresponding linear map of X, then $T^*T\in L(\mathbb{R}^p)$ is self adjoint thus according to the spectral theorem it is diagonalizable w.r.t. some orthonormal basis. Sort T^*T's eigenvalues and their corresponding orthonormal eigenvectors s.t. $\lambda_1>\dots>\lambda_p\geq 0$.

$w = \sum_{i=1}^{p} \langle w, u_i \rangle u_i$ and $T^* T_w = \sum_{i=1}^{p} \lambda_i \langle w, u_i \rangle u_i$. Hence, $\langle Tw, Tw \rangle = \langle w, T^* T_w \rangle = \sum_{i=1}^{p} \lambda_i \langle w, u_i \rangle^2$. In addition, $\|w\|^2 = 1 \rightarrow \langle \sum_{i=1}^{p} \langle w, u_i \rangle u_i, \sum_{i=1}^{p} \langle w, u_i \rangle u_i \rangle = \sum_{i=1}^{p} \langle w, u_i \rangle^2 = 1$. Therefore $\sum_{i=1}^{p} \lambda_i \langle w, u_i \rangle^2 \leq \sum_{i=1}^{p} \lambda_1 \langle w, u_i \rangle^2 = \lambda_1$ and this happens when $\langle w, \lambda_1 \rangle = 1 \rightarrow w = u_1$. Furthermore, $\sum_{i=1}^{p} \lambda_i \langle w, u_i \rangle^2 = \sum_{i=1}^{p} \lambda_1 \langle w, u_i \rangle^2 \rightarrow \sum_{i=1}^{p} (\lambda_i - \lambda_1) \langle w, u_i \rangle^2 = 0$ which is true only if $\langle w, u_i \rangle = 0 \ \forall i = 2, \dots, p$.

(b) $\sum_{i=1}^{n} \operatorname{dist}(x_i, w)^2 = \sum_{i=1}^{n} \langle x_i - \langle w, x_i \rangle w, x_i - \langle w, x_i \rangle w \rangle = \sum_{i=1}^{n} \langle x_i, x_i \rangle - 2 \langle w, x_i \rangle^2 + \langle w, x_i \rangle^2 = \sum_{i=1}^{n} \|x_i\|^2 - \langle w, x_i \rangle^2 = \|X\|_F^2 - \sum_{i=1}^{n} |w^T x_i|^2$.

Therefore $w = u_1$ that maximizes $\sum_{i=1}^{n} |w^T x_i|^2$ also minimizes $\sum_{i=1}^{n} \operatorname{dist}(x_i, w)^2$ and $PC_1^{var} = PC_1^{LS}$.

Question 4

(a) Suppose Y is a random variable with finite expectation and variance. Let $a \in \mathbb{R}$ and define $M(a) = E[(Y-a)^2]$. Then $M(a) = E[Y^2] - 2aE[Y] + a^2$ $\quad \frac{\partial}{\partial a} M(a) = -2E[Y] + 2a = 0 \rightarrow a = E[Y]$. Alternatively, we can treat Y and $a\mathbb{1}$ as vectors in a Hilbert space L^2 with some (Ω, F, P) with the inner product defined as $\langle A, B \rangle = E[AB]$. Thus, we're seeking to minimize $Y - a\mathbb{1}$ by finding the projection of Y onto $\operatorname{span}(\mathbb{1})$. Such a projection $\hat{a}\mathbb{1}$ satisfies $\langle Y - \hat{a}\mathbb{1}, \mathbb{1} \rangle = 0$ $\rightarrow \langle Y, \mathbb{1} \rangle = \hat{a} \langle 1, 1 \rangle \rightarrow \hat{a} = E[Y]$.

(b) Suppose X, Y are random variables with finite expectation and variance. Define $MSE = E[(Y - f(x))^2]$ for some function $f: \mathbb{R} \rightarrow \mathbb{R}$. According to the law of total expactation $MSE = E[E[(Y - f(x))^2 | X]]$. Based on (a), for all x in the support of X $E[(Y - f(x))^2 | X = x]$ is minimized when $\widehat{f(x)} = E(Y | X = x)$, hence, defining $\widehat{f(x)} = E(Y | X = x)$ would minimize the MSE.