Question 1

Proof that:

$$A = \sum_{i=1}^{r} \sigma_i * \mathbf{u_i} * \mathbf{v_i^T} \tag{1}$$

What are we exactly trying to proof?
We're trying to prove that the first k singular vectors provide a linear subspace W which maximizes the squared-sum of the projections of the data onto W.

SVD Theorem:
Computing the best k-dimensional subspace reduces to k applications of the one-dimensional problem.

Scenario for $k = 1$
Lets imagine A is an n by 3 matrix representing n data points x_i in $\mathbb{R}^3$.
Then, each row of A will be the vector representation of each of the points.
Since $k = 1$, the $span\{v\}$ is a line. We want to find the line which is the best fit for the the data, meaning, the sum of the squared projections into that line is minimal.
We could make use of Eq. 4 in matrix notation. Since Av contains as each entry the dot products $x_i v$, **the squared norm of this vector $|Av|^2$ is exactly the sum of the squared lengths of the projections**.
Note also that, simply by Pythagorean, minimizing the squared distances is equivalent to maximize the squared projections.
Therefore, we have rephrased the problem of finding the best subspace of 1D into maximizing $|Av|^2$, which is also equivalent to maximize $|Av|$.

Increasing k
The singular vectors of A are defined recursively as the solutions to these sub-problems.
That is, I'll call v_1 the first singular vector of A, and it is

$$v1 = \arg\max_{v,|v|=1} |Av|$$

Informally speaking, $\sigma_1(A)^2$ represents how much of the data was captured by the first singular vector. Meaning, how close the vectors are to lying on the line spanned by v_1. Larger values imply the approximation is better. *In fact, if all the data points lie on a line, then $\sigma_1(A)^2$ is the sum of the squared norms of the rows of A.*

When extending now to $k = 2$, we are now going to project our data A into a 2D plane V, instead of the previous 1D line. Meaning, the $span\{v_1, v_2\}$ is now a plane.
Because we have already defined v_1, in order to keep creating an orthonormal basis, we consider for v_2 only those that are perpendicular to v_1. Therefore:

$$v2 = \arg\max_{v \perp v_1, |v|=1} |Av| \tag{2}$$

And the SVD theorem implies the subspace spanned by v_1, v_2 is the best 2-dimensional linear approximation to the data. Likewise $\sigma_2(A) = |Av_2|$ is the second singular value. Its squared magnitude tells us how much of the data that was not "captured" by v_1 is captured by v_2. Remember, $\sigma_2(A)^2$ is the sum of the squared projections of the data points into the 2D plane.

Generalizing for k
The k-th singular vector v_k is the one which maximizes $|Av|$ (v being only considered among unit vectors that are perpendicular to $span\{v1, ..., v_{k-1}\}$). The corresponding singular value $\sigma_k(A)$ is the value of the optimization problem.

There are 2 possibilities that can happen after doing this iterative process:

Case 1: The data does not lie in any smaller dimensional subspace (expected)
You would then reach an v_n and there would be no remaining vectors to choose from. The set of v_i are an orthonormal basis of $\mathbb{R}^n$. **This means that the data in A contains a full-rank submatrix**

Case 2: The data actually lie in a smaller k-dimensional subspace (expected)
If this is the case, then the first k singular vectors will span that subspace. This point would be reach during the optimization problem when every perpendicular v has $Av = 0$.

We already have V. What U should look like?
The way he picked v_i to make A diagonal suggests to use Av_i as the columns of U. Then, we define $u_i = Av_i$, the images of the singular vectors under A. Since v_i form an orthonormal basis, $x = \sum_i (x \cdot v_i) v_i$

What exactly A does to any given vector x in terms of its decomposition?
Since v_i form an orthonormal basis, $x = \sum_i (x \cdot v_i) v_i$

$$\begin{aligned} Ax &= A\left(\sum_i (x \cdot v_i) v_i\right) \\ &= \sum_i (x \cdot v_i) A_i v_i \\ &= \sum_i (x \cdot v_i) \sigma_i u_i \end{aligned}$$

That last line can be rewritten as an outer product of two vectors. Therefore:

$$A = \sum_i \sigma_i u_i v_i^T$$