Homework 2 Questions

Q1: Suppose we have three prototype images for Apple, Banana and Orange given as: x_a, x_b, x_o and a test image x_t. This question explores the connection between Cosine similarity C(x, y) and Euclidean distance, E(x, y) between a pair of images x and y.

Each of the image is an N-dimensional vector and $||x_a||_2 = ||x_b||_2 = ||x_o||_2 = ||x_t||_2 = N$

i. (4 marks) Relate the Euclidean distance of the test image with a prototype to its cosine similarity to the same prototype. Derive the formula:

$E(x, x_t) =$

ii. (3 marks) Plot E(x, x_t) vs C (x, x_t)

$iii.$ $(1$ $marks)$ If $C(x_a, x_t) = 0.5$, then E $(x_a, x_t) =$

iv. (2 marks) If cosine similarities of x_t with $x_a, x_b,$ and x_o are 0.5, 0.45 and 0.6 respectively, which of the statement(s) are true:

a. x_a will have the minimum Euclidean distance to x_t

b. x_o will have the minimum Euclidean distance to x_t

c. We can't conclude anything about minimum Euclidean distance to a prototype as it depends on the value of N

d. x_t must be classified as an orange.

Q2. (10 marks) Engineers from American Motors (AMC) are working on a new car (AMX). They are analyzing how the mileage of this car matches up with their other cars and the competition. It is common knowledge that the mileage of a car is dependent on its Engine's displacement or size of the engine in CC and the Weight, W.

They develop an equation to predict mileage as miles per gallon (mpg) using the following regression model where CC is measured in units of 100cc and W in tonne.

$$P = w_o + w_1 CC + w_2 W$$

car name	CC_i	W_i	Actual Mpg (M_i)	Predicted P_i	Error $(E_i = P_i - M_i)$
ford pinto	1	2	25		
amc gremlin	2.3	2.6	19		
amc hornet sportabout (sw)	2.6	3	18		
ford torino 500	2.5	3.3	19		
ford galaxie 500	3.5	4.2	14		

Write the information in a matrix form, write Normal Equations for solving regression problem and obtain the least squares fit for the data. You may use a calculator, but don't use a computer to solve this problem.

Q3. [10 marks] Let A be a square $n \times n$ matrix whose rows are orthonormal. Prove that the columns of A are orthonormal.

Q4. [20 marks] Suppose A is a $n \times n$ matrix with block diagonal structure with k equal size blocks where all entries of the i^{th} block are a_i with $a_1 > a_2 > a_3 \ldots > a_k > 0$. Show that A has exactly k nonzero singular vectors $\boldsymbol{v_1}, \boldsymbol{v_2} \ldots \boldsymbol{v_k}$ where $\boldsymbol{v_i}$ has the value $\sqrt{\frac{k}{n}}$ in the coordinates corresponding to the i^{th} block and 0 elsewhere. In other words, the singular vectors exactly identify the blocks of the diagonal.

What happens if $a_1 = a_2 = a_3 \ldots = a_k > 0$? In the case where the a_i are equal, what is the structure of the set of all possible singular vectors?
Hint: By symmetry, the top singular vector's components must be constant in each block.

Q5. [10 marks] Suppose A is square, but not necessarily invertible and has SVD $A = \Sigma_{i=1}^{r} \sigma_i u_i v_i^T$
Let B = $\Sigma_{i=1}^{r} \frac{1}{\sigma_i} v_i u_i^T$. Show that $BA\,\boldsymbol{x} = \boldsymbol{x}$ for all x in the span of right singular vectors of A.
For this reason, B is sometimes called the pseudo inverse of A and can play the role of A^{-1} in many applications.

Q6. [10 marks]

a. For any matrix A, show that $\sigma_k \leq \frac{||A||_F}{\sqrt{k}}$

b. Prove that there exists a matrix B for rank at most k such that $\left||A - B|\right|_2 \leq \frac{||A||_F}{\sqrt{k}}$

Where the 2 norm of a matrix is defined as follows:

$$\left||A|\right|_2 = \max\left||Ax|\right|_2 \; s.t. \left||x|\right| = 1$$

Q7. [20 marks] Fast Approximate Matrix Vector Multiplication Algorithm

Suppose an n×d matrix A is given and you are allowed to preprocess A.
Then you are given a large number of d-dimensional vectors $x_1, x_2, \ldots . x_m$ and for each of these vectors you must find the vector Ax_j approximately, in the sense that you must find a vector y_j satisfying $|y_j - Ax_j| \leq \epsilon\left||A|\right|_F |x_j|$.
Here ϵ>0 is a given error bound.
Describe an algorithm that accomplishes this in time $O(\frac{n+d}{\epsilon^2})$ per x_j not counting the preprocessing time.

Hint: Use Q6

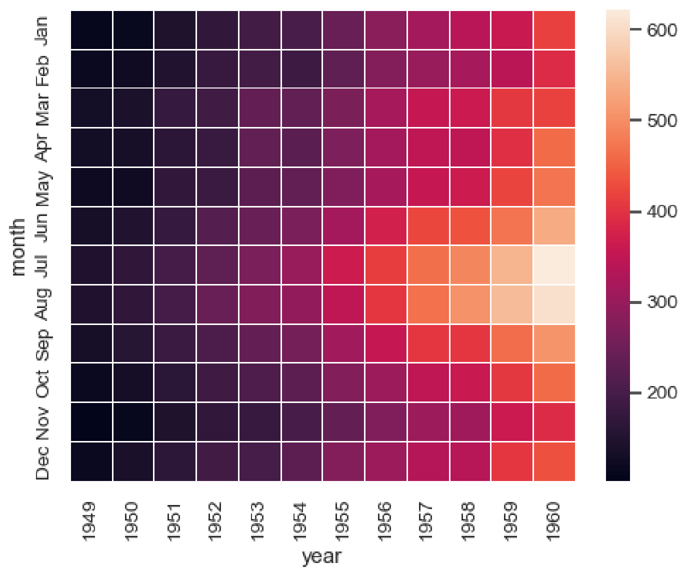
Q8. [10 marks] Find the closest rank-1 approximation to the following matrices (in Frobenius norm)

a. $\boldsymbol{A} = \begin{bmatrix} 0 & 0 & 3 \\ 0 & 2 & 0 \\ 1 & 0 & 0 \end{bmatrix}$

b. $\boldsymbol{A} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$

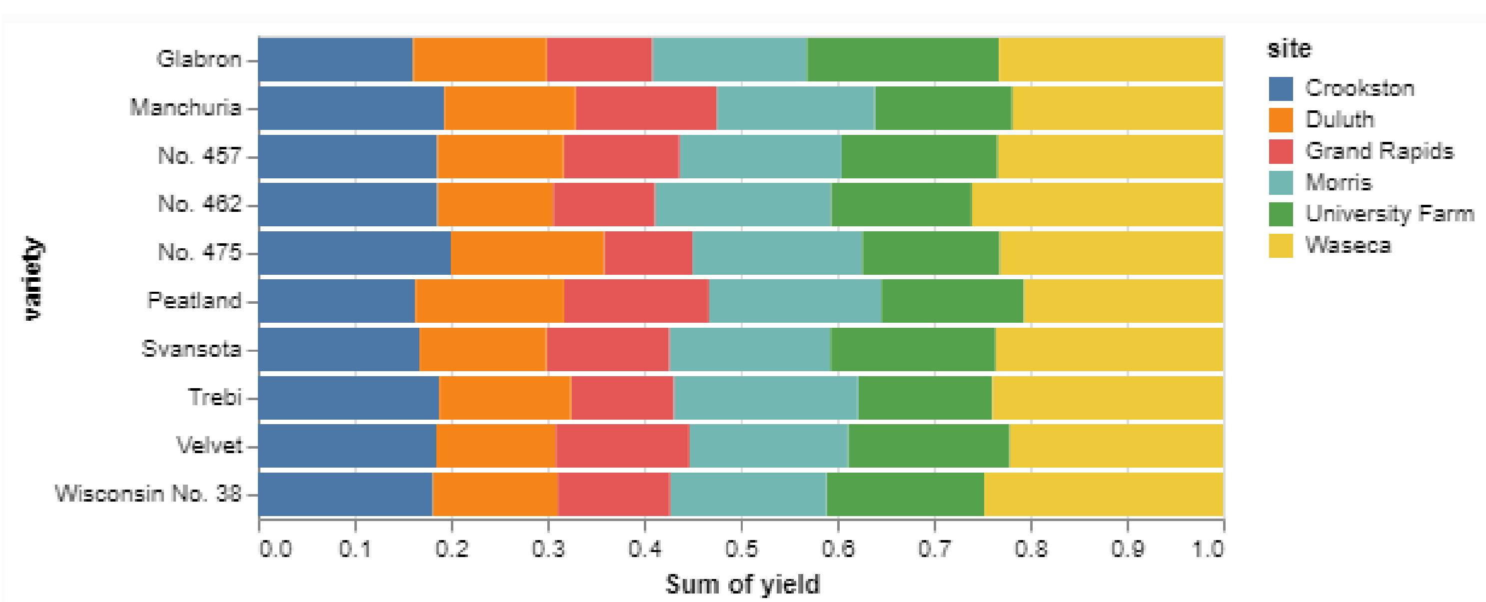
Extra Practice Questions from a past exam. DO NOT SUBMIT THEIR SOLUTIONS

Q3a. [5 marks] This is a heatmap plot of passengers who took flights during 1949 to 1960.

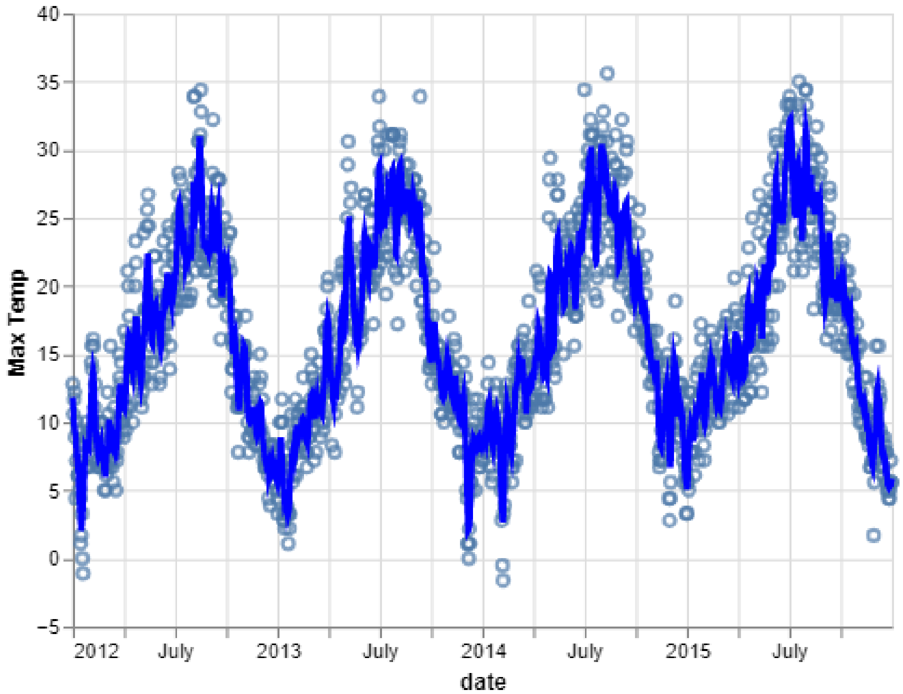


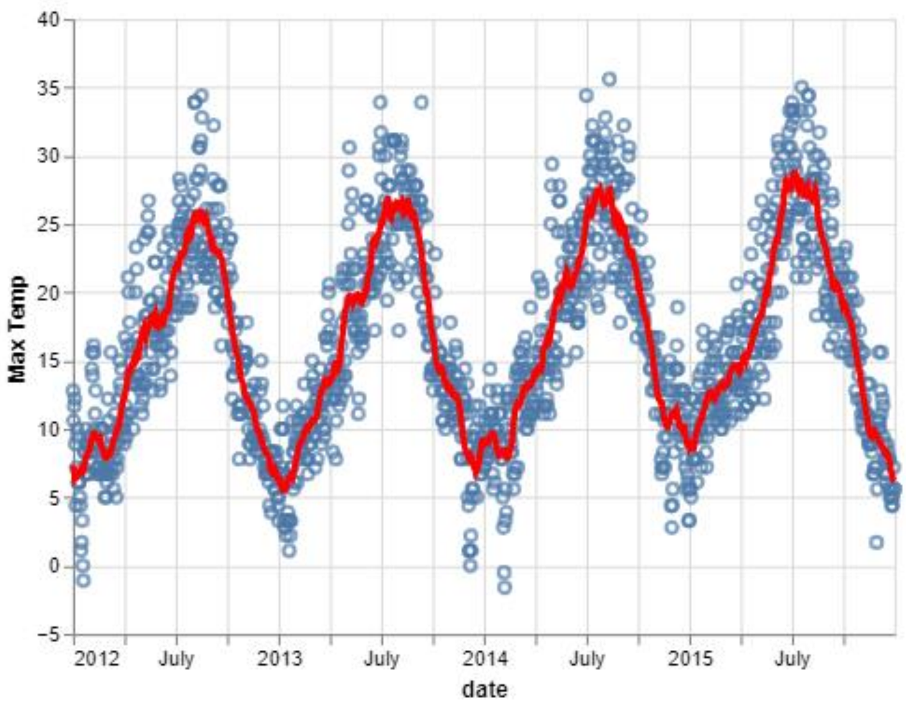
Identify the busiest month in this whole period.

Q3b. [5 marks] This is a normalized stacked bar chart using data which contains crop yields over different regions and different years in the 1930s. Which site has the least production of No. 475 variety?



Q 4 [10 marks] The max temperature in the city of Seattle is plotted here as a scatter-plot with Rolling mean. There are two plots here, one with rolling mean window of 30 days and another with window of 5 days. Identify them and write what they each represent?





Q5. [10 marks] This example is a fully developed line chart that uses a window transformation. Write a brief interpretation of this chart (no more than 50 words).