CS215 Assignment-1

Abhi Jain - 23b0903
Anushka Singhal - 23b0928
Sabil Ahmad - 23b1057

August 2024

Contents

1 Let's Gamble

$$P(\text{A will have more wins than B}) = \frac{\sum_{0\le i<j\le n+1} \binom{n}{r}\binom{n+1}{j}}{2^n \dot{2}^{n+1}}$$

$$P = \frac{\sum_{i=0}^{n} \binom{n}{i}\binom{n+1}{i+1} + \sum_{i=0}^{n-1} \binom{n}{i}\binom{n+1}{i+2} + \ldots + \sum_{i=0}^{1} \binom{n}{i}\binom{n+1}{i+n} + \binom{n}{0}\binom{n+1}{n+1}}{2^{2n+1}}$$

$$P = \frac{\sum_{i=0}^{n} \binom{n}{n-i}\binom{n+1}{i+1} + \sum_{i=0}^{n-1} \binom{n}{n-i}\binom{n+1}{i+2} + \ldots + \sum_{i=0}^{1} \binom{n}{n-i}\binom{n+1}{i+n} + \binom{n}{n-0}\binom{n+1}{n+1}}{2^{2n+1}}$$

Using Vandermonde identity, $\binom{m+n}{r} = \sum_{k=0}^{r} \binom{m}{k}\binom{n}{r-k}$

$$P = \frac{\binom{2n+1}{n+1} + \binom{2n+1}{n+2} + \ldots + \binom{2n+1}{2n+1}}{2^{2n+1}}$$

which can be rewritten as

$$P = \frac{\binom{2n+1}{n} + \binom{2n+1}{n-1} + \ldots + \binom{2n+1}{0}}{2^{2n+1}}$$

Adding above two equations,

$$2P = \frac{\binom{2n+1}{0} + \binom{2n+1}{1} + \ldots + \binom{2n+1}{2n} + \binom{2n+1}{2n+1}}{2^{2n+1}}$$

$$\implies 2P = \frac{2^{2n+1}}{2^{2n+1}}$$

$$\implies P = \frac{1}{2}$$

Alternate Solution

Let's first consider the case when both have just thrown and checked n dice throws only and A is yet to throw his $(n+1)^{th}$ dice.

Let $P(A, I)$ denote the event that A wins and both have thrown n dice, while $P(A, II)$ denote the event that A wins, A has thrown $n+1$ dice and B has thrown n dice. We have 3 mutually exclusive and exhaustive events i.e., either A wins, a draw or B wins. Therefore,

$$P(A, I) + P(\text{tie}, I) + P(B, I) = 1$$

By symmetry for this case, we have $P(A, I) = P(B, I)$. Hence,

$$P(A, I) + \frac{1}{2}P(\text{tie}, I) = \frac{1}{2} \tag{1}$$

Now, A is making the $(n+1)^{th}$ throw. If A has already won, then this throw won't affect the result. If it was a tie then the probability that A wins is the

probability he gets the prime in this throw. If B has won the game before, then this outcome can only turn into a tie or remain B's win. therefore,

$$P(A, II) = P(A, I) + P(\text{tie}, I) \cdot P(\text{prime on a dice throw})$$

$$\implies P(A, II) = P(A, I) + \frac{1}{2}P(\text{tie}, I)$$

This can be evaluated using the 1 as

$$P(A \text{ wins}) = P(A, II) = \frac{1}{2}$$

2 Two Trading Teams

Let the probability of winning from A be 'a' and that from B be 'b'.

Now we know that B is a better trader than A, so it will be more difficult to win from B than from A, so the probability of winning from B will be less than winning from A, therefore $a > b$.

Now we will win when we win two consecutive sets, so for

- $A - B - A$: There will be two cases, one if it wins the first two sets, and the second if it looses the first set and wins the last two sets. Let the probability of the two cases be p_1 and p_2 respectively.

$$p_1 = a \times b$$

$$p_2 = (1 - a) \times b \times a$$

$$P(A - B - A) = p_1 + p_2 = ab + (1 - a)ab$$

$$P(A - B - A) = ab(2 - a)$$

- $B - A - B$: Again there will be the same two cases as above, so let their probabilities be p_3 and p_4 respectively.

$$p_1 = b \times a$$

$$p_2 = (1 - b) \times a \times b$$

$$P(B - A - B) = p_1 + p_2 = ab + (1 - b)ab$$

$$P(B - A - B) = ab(2 - b)$$

Now as,

$$a > b \implies -a < -b$$

$$\implies 2 - a < 2 - b$$

$$\implies ab(2 - a) < ab(2 - b)$$

$$\implies P(A - B - A) < P(B - A - B)$$

So, we will choose the option B-A-B.

3 Random Variables

3.1 Part 1

Let Event A = $Q_1 < q_1$, and B = $Q_2 < q_2$
Given $P(A) \geq 1 - P_1$ and $P(B) \geq 1 - P_2$
For the event C = $Q1.Q2 < q1.q2$,
Clearly , $A \cap B \subset C$

$$A \cap B \subset C \implies P(A \cap B) \leq P(C)$$

$$P(A \cap B) = P(A) + P(B) - P(A \cup B)$$
$$\implies P(A) + P(B) - P(A \cup B) \leq P(C)$$

Now, since

$$P(A) \geq 1 - P_1$$
$$P(B) \geq 1 - P_2$$
$$-P(A \cup B) \geq -1$$

On adding all the above three equations we get

$$P(A) + P(B) - P(A \cup B) \geq 1 - P_1 - P_2$$
$$\implies P(C) \geq P(A) + P(B) - P(A \cup B) \geq 1 - P_1 - P_2$$
$$\implies P(C) \geq P(A \cap B) \geq 1 - (P_1 + P_2)$$
$$\implies P(C) \geq 1 - (P_1 + P_2)$$

Hence, Proved.

3.2 Part 2

Let's assume there exists an j such that $|x_j - \mu| > \sigma(n-1)$
The standard deviation

$$\sigma^2 = \sum_{i=1}^{n} \frac{|x_i - \mu|^2}{n-1}$$

$$\sigma^2 = \frac{|x_j - \mu|^2}{n-1} + \sum_{i=1, i \neq j}^{n} \frac{|x_i - \mu|^2}{n-1}$$

Since for $x_j, |x_j - \mu| > \sigma(n-1)$ clearly,

$$\frac{|x_j - \mu|^2}{n-1} > \sigma^2$$

$$\implies \frac{|x_j - \mu|^2}{n-1} + \sum_{i=1, i\neq j}^{n} \frac{|x_i - \mu|^2}{n-1} > \sigma^2$$

But, the term on the left is the standard deviation itself,

$$\implies \sigma^2 > \sigma^2$$

Hence, we have a contradiction thus there can't exist a j such that $|x_j - \mu| > \sigma(n-1)$ and for all j $|x_j - \mu| \leq \sigma(n-1)$

Proven Inequality: The bound $\sigma\sqrt{n-1}$ is definitely tighter, particularly when the data is concentrated around the mean. This inequality gives a concrete limit on how far any single data point can be from the mean, which is particularly useful in deterministic settings or when analyzing small datasets.

Chebyshev's Inequality: Chebyshev's bound is generally looser because it applies to all distributions with finite variance. It is useful when dealing with random variables where the distribution is unknown or non-normal, as it provides a worst-case scenario bound.

4 Staff Assistant

4.1 Part (a)

Given,

- E = Event that we hire the best assistant
- E_i = Event that i^{th} candidate is the best assistant and we hire him/her

The two sub-events that form the event E_i are independent. Hence, we can write $Pr(E_i)$ as:

$$Pr(E_i) = Pr(\text{Best candidate is the } i^{th}\text{candidate}) \cdot Pr(\text{He/She has been hired})$$

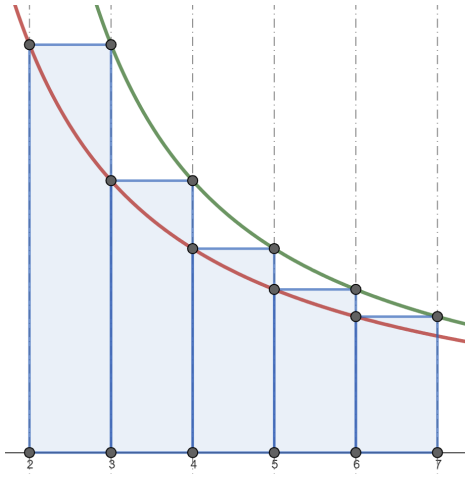
Now, probability that best candidate is i^{th} one is simply $\frac{1}{n}$. The probability that he/she has been hired is the same as the probability that the maximum score of the first $i-1$ candidates occurs in the first m candidates otherwise someone from $m+1$ to $i-1$ candidate would have been hired.

$$\implies Pr(E_i) = \begin{cases} \frac{1}{n} \cdot \frac{m}{i-1}, & i > m \\ 0, & i \leq m \end{cases}$$

Now, $Pr(E)$ can be simply evaluated as:

$$Pr(E) = \sum_{i=1}^{n} Pr(E_i) = \sum_{i=m+1}^{n} Pr(E_i)$$

$$\implies Pr(E) = \frac{m}{n} \cdot \sum_{i=m+1}^{n} \frac{1}{i-1} \tag{2}$$



4.2 Part (b)

Now, we approximate $\sum_{i=m+1}^{n} \frac{1}{i-1}$ using Riemann's integrations.

$$\sum_{i=m+1}^{n} \frac{1}{i-1} = \frac{1}{m} + \frac{1}{m+1} + \cdots + \frac{1}{n-1}$$

$$\int_m^n \frac{1}{i} di \leq \sum_{i=m+1}^{n} \frac{1}{i-1} \leq \int_m^n \frac{1}{i-1} di$$

$$\implies \log(n) - \log(m) \leq \sum_{i=m+1}^{n} \frac{1}{i-1} \leq \log(n-1) - \log(m-1)$$

$$\implies \frac{m}{n}\left(\log(n) - \log(m)\right) \leq Pr(E) \leq \frac{m}{n}\left(\log(n-1) - \log(m-1)\right)$$

Hence, proved.

4.3 Part (c)

Let,

$$f(m) = \frac{m}{n}(\log(n) - \log(m))$$

$$\implies \frac{df}{dm} = \frac{1}{n}(\log(n) - 1 - \log(m))$$

$$\implies \frac{d^2 f}{dm^2} = \frac{-1}{m \cdot n}$$

Now,

$$\frac{df}{dm} = 0 \implies m = \frac{n}{e} \implies \frac{d^2 f}{dm^2} < 0 \left(\text{for } m = \frac{n}{e}\right)$$

Therefore, by second derivative test $f(m)$ is maximum at $m = \frac{n}{e}$

Hence, for $m = \frac{n}{e}$

$$Pr(E) \geq \frac{m}{n}(\log(n) - \log(m)) \implies Pr(E) \geq \frac{1}{e}$$

5 Free Trade

Let's calculate the probability of getting a free trade when there are k people ahead of us, clearly $k \leq 200$ else by PHP two people ahead of us would have the same ID and the latter of the two would get the free trade.
The probability of which is that among the k people ahead of us one of them has the same ID as us and the rest k people all have different IDs, none of them the same as us, which after evaluation comes out to be

$$P_k = \left(\frac{k}{200}\right) \cdot \left(\frac{199!}{(200-k)!(200)^{k-1}}\right)$$

For a maxima

$$P_k \geq P_{k-1} \; and \; P_k \geq P_{k+1}$$

Solving the inequalities we get a range on k and the integer value comes out to be k = 14 .
Thus when there are 14 people ahead of us or at the **15th position** the chances of getting a free trade are maximum.

6 Update Functions

The derivation for the formulas and algorithms are as below. The relevant codes have been added into the submission folder.

6.1 Update Mean

$$\bar{x}_{\text{old}} = \frac{1}{n}\left(\sum_{i=1}^{n} A_i\right)$$

$$\bar{x}_{\text{new}} = \frac{1}{n+1}\left(\sum_{i=1}^{n} A_i + A_{\text{newValue}}\right)$$

$$\implies \bar{x}_{\text{new}} = \frac{n \cdot \bar{x}_{\text{old}} + A_{\text{newValue}}}{n+1}$$

6.2 Update Median

For this, we assume that the array A is sorted to have a $O(1)$ time complexity algorithm to update the median. If the array is unsorted we will have to loop through the array which isn't allowed.

Considering indexing of array A from 0 to $n-1$

- **n = initial size is even**: Then,

$$\text{newMedian} = \begin{cases} A[\frac{n}{2} - 1], & \text{newValue} <= A[\frac{n}{2} - 1] \\ A[\frac{n}{2}], & \text{newValue} >= A[\frac{n}{2}] \\ \text{newValue}, & \text{else} \end{cases}$$

- **n = initial size is odd**: Then,

$$\text{newMedian} = \begin{cases} \frac{1}{2} \cdot \left(A[\frac{n-1}{2} - 1] + \text{oldMedian}\right), & \text{newValue} <= A[\frac{n-1}{2} - 1] \\ \frac{1}{2} \cdot \left(A[\frac{n-1}{2} + 1] + \text{oldMedian}\right), & \text{newValue} >= A[\frac{n-1}{2} + 1] \\ \frac{1}{2} \cdot (\text{newValue} + \text{oldMedian}), & \text{else} \end{cases}$$

6.3 Update Standard Deviation

$$\sigma^2_{\text{old}} = \frac{1}{n-1} \left(\sum_{i=1}^{n} (A_i)^2 - n \cdot \bar{x}^2_{\text{old}} \right)$$

$$\implies \sum_{i=1}^{n} (A_i)^2 = (n-1) \cdot \sigma^2_{\text{old}} + n \cdot \bar{x}^2_{\text{old}}$$

$$\sigma^2_{\text{new}} = \frac{1}{n} \left(\sum_{i=1}^{n} A_i^2 + A^2_{\text{newValue}} - (n+1) \cdot \bar{x}^2_{\text{new}} \right)$$

$$\implies \sigma^2_{\text{new}} = \frac{1}{n} \left((n-1) \cdot \sigma^2_{\text{old}} + n \cdot \bar{x}^2_{\text{old}} + A^2_{\text{newValue}} - (n+1) \cdot \bar{x}^2_{\text{new}} \right)$$

6.4 Histogram

The height of the bar for the bin/interval in which the new value lies will be increased by 1.

7 Plots

7.1 Violin Plot

The plot is mainly descriptive and can be used to compare distribution in different groups. It combines the features of a box plot and a density plot for a detailed view of distribution. Based on the density curve (Kernel Density Estimate), it is possible to compare the height of the peak, the depth of the valley and the length of the tail of different groups of the variables to see if they are centred by the same distribution or not. Furthermore, within each violin, a box plot may be presented, which demonstrates the median along with the quartiles. Although, violin plot is not as common as other plots are, because of the extra complications it involves.

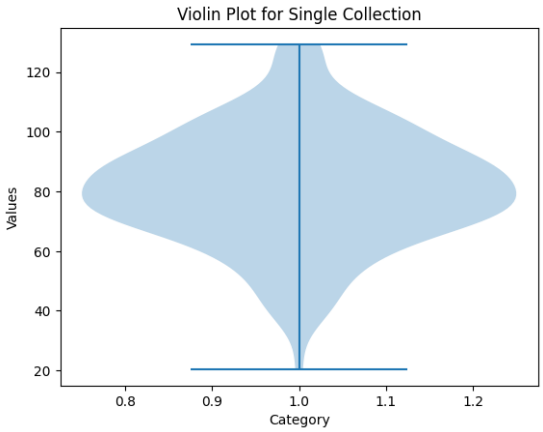


Figure 1: Violin plot

7.2 Pareto Chart

The Pareto Chart is basically a bar chart with a line graph. The vertical bars indicate the individual values in the descending manner and the line graph portrays the cumulative percentage of the above values. This two-way analysis is used for identifying and prioritizing issues, causes, or factors that have the greatest impact within a dataset. It follows the **Pareto principle (the 80-20 rule)** which states that 80% of problems often come from 20% of causes. The benefit of having data in descending order of value or importance is that is allows the decision makers to allocate their focus on the most influential factors of a problem. Therefore it is very helpful when addressing a particular issue or problem. The combination of bars and graphs simplifies pattern recognition and relationship management.

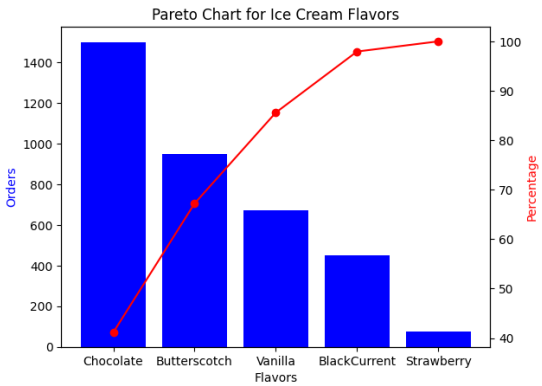


Figure 2: Pareto Chart

7.3 Coxcomb Chart

Coxcomb charts are also known by names such as rose charts, polar area diagrams, and these are an improved version of the pie charts where the data is represented by the area of the sectors rather than angles. While the angles remain constant, the radii of the slices vary based on the data values. These charts became most famous when Florence Nightingale used it to show causes of death regarding soldiers during the Crimean war in the 19th century. Coxcomb charts are more suitable when dealing with data proportions for the different categories, as the segment length (or area) eases comparison of values. They best apply in cases where the main information to present is proportional and where there are different categories or a time frame to follow.

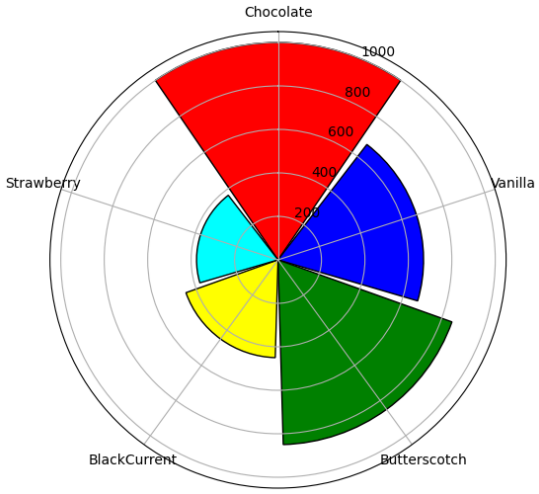


Figure 3: Coxcomb Chart

7.4 Waterfall Plot

Waterfall plots, also known as bridge charts, visualize how a starting value transitions to an ending value by highlighting intermediate steps. They are commonly used in business and finance to track sequential positive and negative changes that impact a final total, such as revenue, expenses, or profits. In this chart, since different colors are used to categorize gains and the loses that fall in the different factors, there is ease in identifying how each factor has contributed to the final figure. Its structure is very logical and demonstrates the process step by step which is helpful in illustrating the visual data and guiding the audience through the data presented.

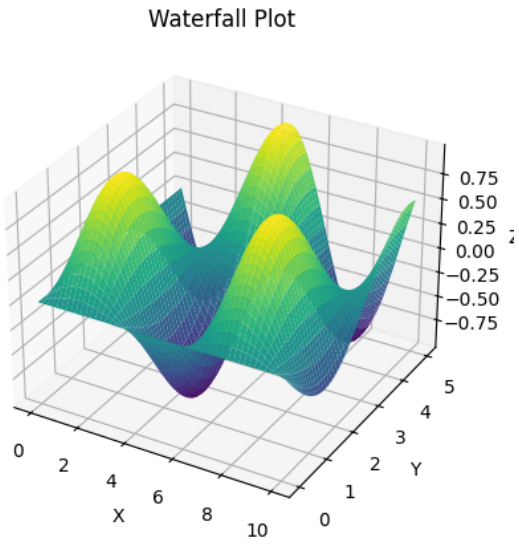


Figure 4: Waterfall plot

8 Monalisa

Here is our original image of the Mona Lisa which we shifted by different pixel values to calculate the correlation factor between it and the original image.

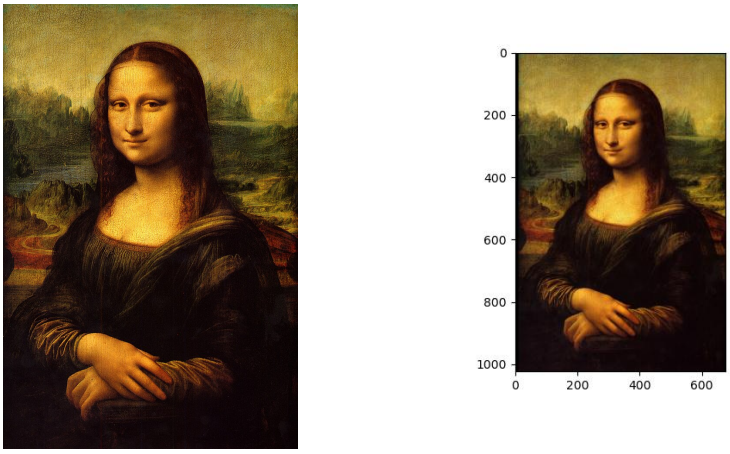


Figure 5: Original Image compared with an image shifted by 10 pixels left

On doing this for shift ranges from -10 to 10 we obtain the following graph

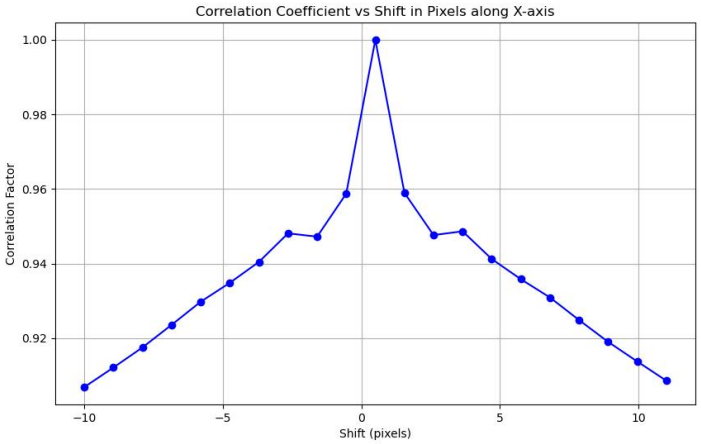


Figure 6: Correlation Coefficient vs Shift in Pixels

Now we generate the normalized histogram for the image for R , G , B and the gray-scale Channels.

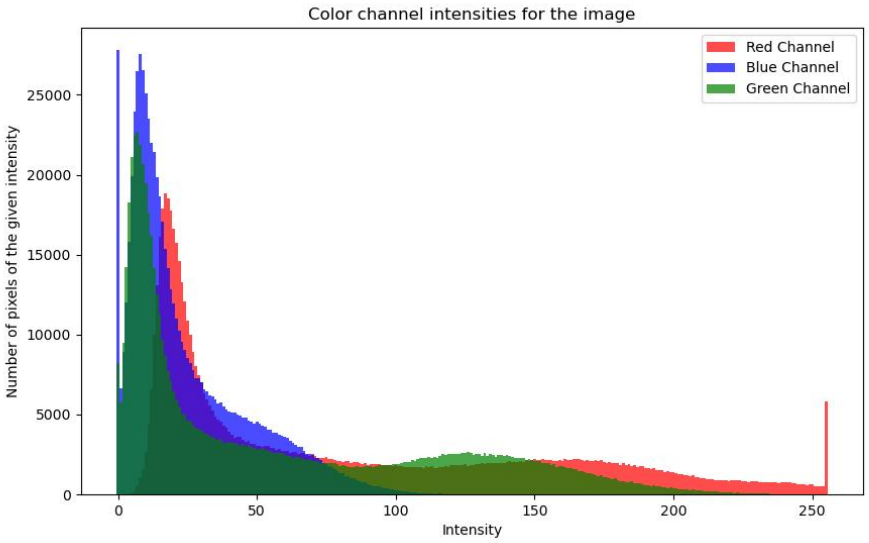


Figure 7: Histogram for Different Color Intensities

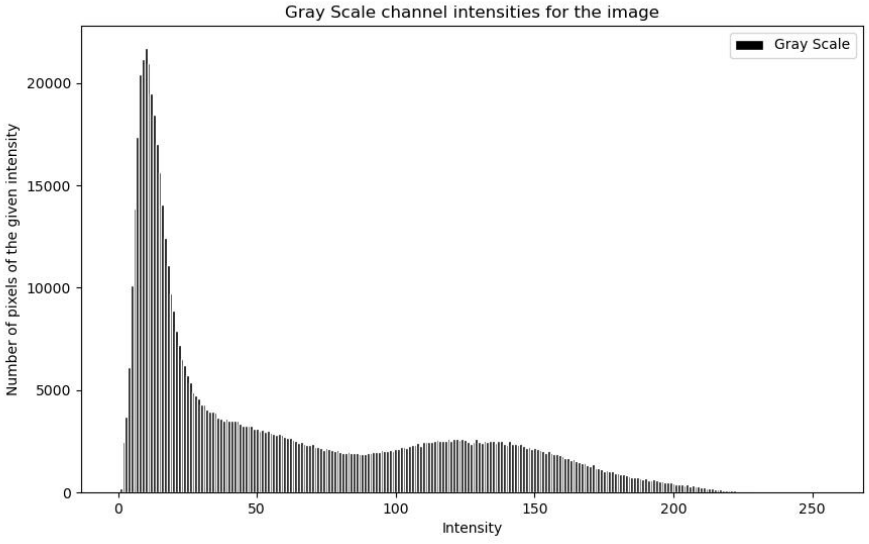


Figure 8: Histogram for gray-scale