Optimal Histogram Construction
Meysam Moradi Modiryan

1 Introduction

The introduction of histograms was done by Pearson [4] for the first time. He proposed using them as a tool to investigate the distribution of a given dataset approximately. In this problem, the aim is to divide the range of data into a number of intervals which are called *bins*. It is important to select the points in which the non-overlapping intervals have to break. Therefore, we can call the bin width as the most important parameter that must be determined in an appropriate way in order to obtain the fundamental structure of the data. The *over smoothing* and *under smoothing* of the histograms which mean providing insufficient or plenty details respectively, depend heavily on the precise estimation of the bin width [7].

Since histograms could be applicable in different areas, they have gained the concentration of many studies in the literature. In [3], the authors define histograms as non-parametric estimators for the density which are utilized to present data concisely. Stating the importance of converging to the precise density function, the authors propose a method based on the integrated mean squared error minimization to capture the optimal histogram. The authors in [5] investigate the role of Bayesian Probability Theory to develop an algorithm in which histograms are treated as piecewise-constant model of the probability density. In addition, aiming of reducing the subjectivity while incorporating the *smoothing* parameter, the authors of [2] use Akaike Information Criteria (AIC) [1] to optimize the histograms.

In general, histograms are simple and beneficial tools for data visualization. The rules of determining the bins' number and size seem not to be complicated, however, it is not yet obvious that which rule lead to the optimal histogram. Considering the objective of minimizing a conic representable distance measure like the Wassertein metric [6] and/or Kullback-Leibler divergence (KL-divergence) [8], our aim is to develop a mixed integer conic programming model to determine the number and size of the bins. We will test the accuracy of the model on the real data of students' grades to determine the associated letter grade optimally.

2 Wasserstein Distance Metric

If P and Q are the empirical distribution of two datasets defined on the probability space $\mathcal{X}$, then the Wasserstein distance takes a function of the order statistics:

$$W_p(P,Q) = \left(\sum_{x\in\mathcal{X}} \|P(x) - Q(x)\|^p \right)^{1/p}$$

If we get $p = 1$, therefore, minimizing the Wasserstein distance will be as follows.

$$\min_{Q} \sum_{x\in\mathcal{X}} \|P(x) - Q(x)\|)$$

With its equivalent optimization problem as:

$$\begin{aligned} \min_{t,Q} & \sum_{x\in\mathcal{X}} t(x) \\ \text{s.t.} \quad & \|P(x) - Q(x)\|^2 \le t(x) \end{aligned}$$

This problem can be written as a Second Order Conic Programming (SOCP) problem as follows.

$$\min_{t,Q} \sum_{x\in\mathcal{X}} t(x)$$
$$\text{s.t.} \begin{bmatrix} P(x) - Q(x) \\ t(x) \end{bmatrix} \in \mathcal{L}^2$$

Where $\mathcal{L}^3$ is the Lorentz cone of order 3.

We have constructed the conic programming problem of minimizing the Wasserstein distance metric so far. In the next section, we are going to add some constraints to this problem in order to optimize the histogram construction problem.

3 Optimal Histogram Construction

The assumptions of this problem as well as the majority of the constraints are inspired from the master's thesis of Sahand Asgharieh Ahari titled "Mixed-Integer Exponential Cone Programming in Action: Sparse Logistic Regression and Optimal Histogram Construction". The aim of this project is to apply the Wasserstein metric as a second-order cone representable distance measure for converting numerical grades to letter grades.

3.1 Assumptions

The assumptions one may take into account to construct the optimal histogram problem as a Mixed-Integer Second Order Conic Programming problem could be:

- At least one observation should be dedicated to each bin.
- Each observation must be allocated to exactly one bin.
- If observation θ_i is dedicated to bin j, then the next observation θ_{i+1} should be dedicated to either bin j or $j+1$.

3.2 Decision Variables

We need to define a set of decision variables as follows.

$$x_{ij} = \begin{cases} 1 & \text{if observation } i \text{ is allocated to bin } j, \\ 0 & \text{O.W.} \end{cases}$$

$$q_i = \text{ probability of } \theta_i \text{ in estimated distribution}$$

$$Q_j = \text{ probability of bin j in estimated distribution}$$

As an instance, let us define the vector of $p = [0.1, 0.2, 0.4, 0.3]$ to represent the observations θ_i's probabilities. Assume $k = 2$ and the model associates the first two observations to the first bin and the remainder to the second bin, then the value of Q_1 and Q_2 are 0.3 and 0.7, respectively. Consequently, the values of q_1 and q_2 will be equal to 0.15 while those of q_3 and q_4 will be equal to 0.35.

3.3 Mathematical Model

Our MISOCP model for the optimal histogram construction problem is given as:

$$\min_{q,t,x,Q} \sum_{i=1}^{m} t_i \tag{1}$$

$$\text{s.t.} \begin{bmatrix} p_i - q_i \\ t_i \end{bmatrix} \in \mathcal{L}^2 \qquad \forall i \tag{2}$$

$$\sum_{i=1}^{m} x_{ij} \geq 1 \qquad \forall j \tag{3}$$

$$\sum_{j=1}^{n} x_{ij} = 1 \qquad \forall i \tag{4}$$

$$x_{ij} \leq x_{i+1,j} + x_{i+1,j+1} \qquad \forall i,j \tag{5}$$

$$\sum_{j'=1}^{j} x_{ij'} + \sum_{j''=j+2}^{n} x_{i+1,j''} \leq 1 \qquad \forall i,j \tag{6}$$

$$Q_j = \sum_{i'=1}^{m} p_{i'} x_{i'j} \qquad \forall j \tag{7}$$

$$\frac{Q_j}{\sum_{i'=1}^{m} x_{i'j}} - (1 - x_{ij}) \leq q_i \leq \frac{Q_j}{\sum_{i'=1}^{m} x_{i'j}} + (1 - x_{ij}) \qquad \forall i,j \tag{8}$$

Where $i = 1, \ldots, m, i' = 1, \ldots, m$, and $j = 1, \ldots, n$. Constraint (3) ensures that at least one observation should be dedicated to each bin. To allocate each observation to exactly one bin we have (4). Constraints (5) and (6) are equivalently certifying that if observation θ_i is dedicated to bin j, then the next observation θ_{i+1} should be dedicated to either bin j or $j+1$. In order to make sure that the constraint (5) is working precisely for all the observations and bins pairs, we have to break it as follows:

$$\begin{aligned} & x_{ij} \leq x_{i+1,j} + x_{i+1,j+1} && i = 1, \ldots, m-1, j = 1, \ldots, n-1 \\ & x_{ij} \leq x_{i+1,j} && i = 1, \ldots, m-1, j = n \\ & x_{i-1,j} \leq x_{ij} && i = m.j = n \end{aligned}$$

Constraints (7) and (8) determine the distribution of the observation is each bin. The last constraint is nonlinear and we will linearize it. First we will omit the fractional expressions and then define new variables to deal with the nonlinearity. We can write the constraint (8) as follows.

$$Q_j - (1 - x_{ij}) \sum_{i'=1}^{m} x_{i'j} \leq q_i \sum_{i'=1}^{m} x_{i'j} \leq Q_j + (1 - x_{ij}) \sum_{i'=1}^{m} x_{i'j} \qquad \forall i,j$$

Therefore, we have two nonlinear expressions to linearize by defining new variables.

Expression 1:

$$(1 - x_{ij}) \sum_{i'=1}^{m} x_{i'j}$$

Let us define new variables $\alpha_{ii'j}$ and u_{ij} as follows.

$$\alpha_{ii'j} = (1 - x_{ij})x_{i'j} \qquad \forall i, i', j \tag{9}$$

$$u_{ij} = \sum_{i'=1}^{m} \alpha_{ii'j} \qquad \forall i, j \tag{10}$$

$$0 \leq \alpha_{ii'j} \leq x_{i'j} \qquad \forall i, i', j \tag{11}$$

$$(1 - x_{ij}) - (1 - x_{i'j}) \leq \alpha_{ii'j} \leq (1 - x_{ij}) \qquad \forall i, i', j \tag{12}$$

Expression 2:

$$q_i \sum_{i'=1}^{m} x_{i'j}$$

Analogously, we define new variables $\beta_{ii'j}$ and v_{ij} as follows.

$$\beta_{ii'j} = q_i x_{i'j} \qquad \forall i, i', j \tag{13}$$

$$v_{ij} = \sum_{i'=1}^{m} \beta_{ii'j} \qquad \forall i, j \tag{14}$$

$$0 \leq \beta_{ii'j} \leq x_{i'j} \qquad \forall i, i', j \tag{15}$$

$$q_i - (1 - x_{i'j}) \leq \beta_{ii'j} \leq q_i \qquad \forall i, i', j \tag{16}$$

Substituting these new defined linear constraints in the model yields the final MISOCP model for the optimal histogram construction problem as follows.

$$\begin{aligned} \min_{q,t,x,Q} & \sum_{i=1}^{m} t_i \\ \text{s.t.} & (2) - (7), \\ & (10) - (12), \\ & (14) - (16), \\ & Q_j - u_{ij} \leq v_{ij} \leq Q_j + u_{ij} \qquad \forall i, j \end{aligned}$$

We will present the computational results of utilizing this model for our experiments in the next section.

4 Computational Results

In this section we are going to present the results of utilizing the proposed model to construct the optimal histograms of the letter grades. At the end of each semester, the total grade of a student for a course is calculated which is out of 100. The instructors have to enter the associated letter grades for the numerical grade of the students in the related portal. The process of converting numerical grades to letter grades is usually performed manually since the grades are not distributed uniformly between 0 and 100. Otherwise, it was just a simple *if* condition (e.g. if the grade ≥ 85 then assign A). As the number of students matters in this manual process and there is the possibility of human errors in converting the grades, we try to do this process by our mathematical MISOCP model.

We have the grades of the students for the Operations Research course for four semesters which have to be assigned to 10 bins as we have 10 letter grades: $A, A^-, B^+, B, B^-, C^+, C, C^-, D^+, D$. As a rule, numerical grades less than 40 are failed with the F letter grade therefore, we do not take them into account of the experiment and have deleted those grades. The assignment will be done for the grades between or equal to 40 and 100.

4.1 Data Visualization

In this part we are going to give an insight of the data by visualizing them. In figures (1)-(4) we can see the numerical grades of the students and their assigned letter grades by their instructor as well as the frequency of each numerical grade. The distribution of the probabilities of each individual letter grade which looks like the frequency of the grades is provided but since the probability is a small number, it is not clear from the figures. The comparisons between the data and the results of the model will be covered in the next section.

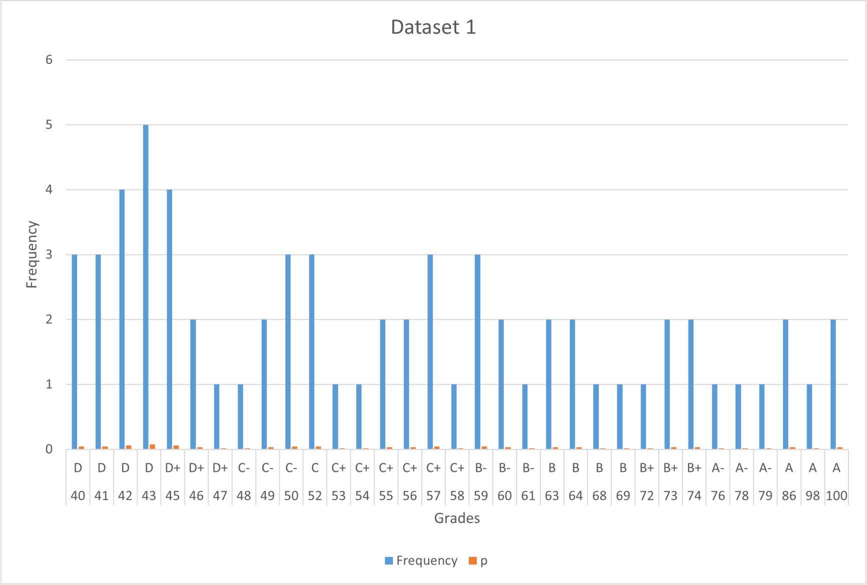


Figure 1: Numerical and Letter Grades of Dataset 1 (Spring 2019)

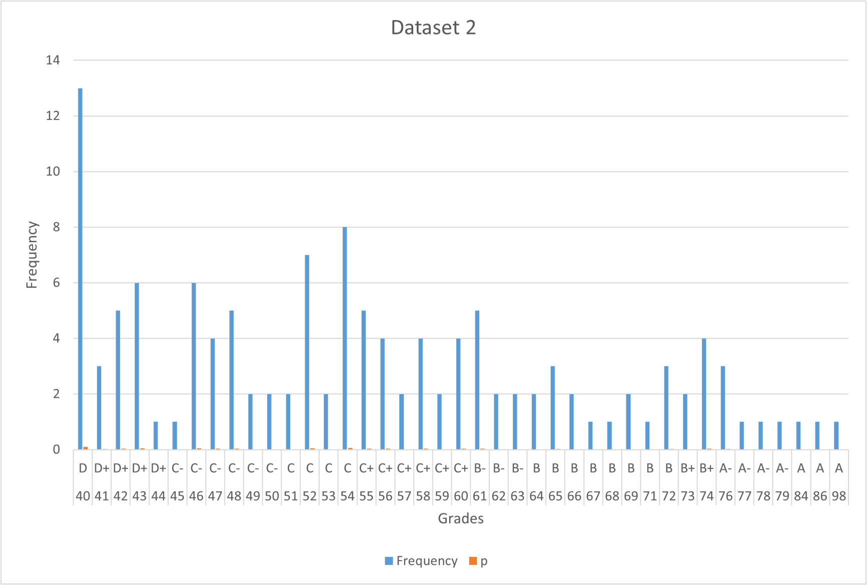


Figure 2: Numerical and Letter Grades of Dataset 2 (Spring 2020)

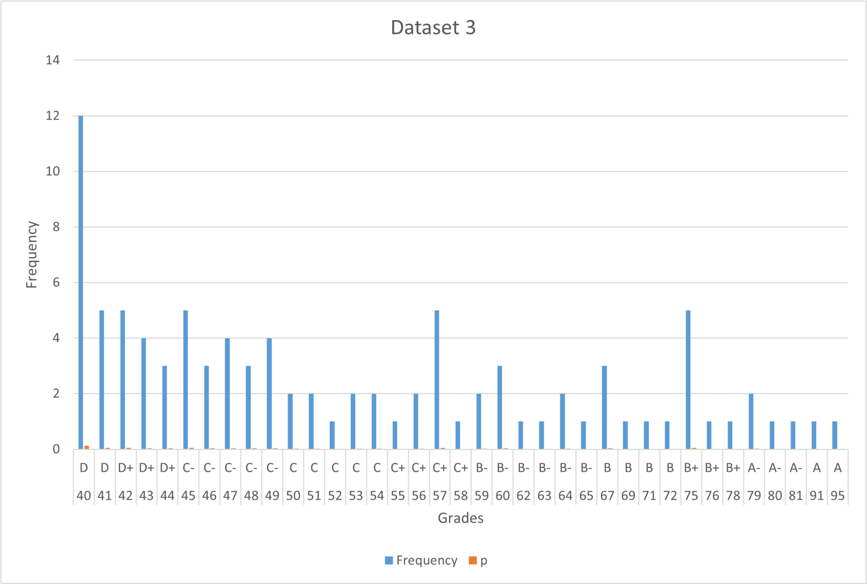


Figure 3: Numerical and Letter Grades of Dataset 3 (Fall 2020)

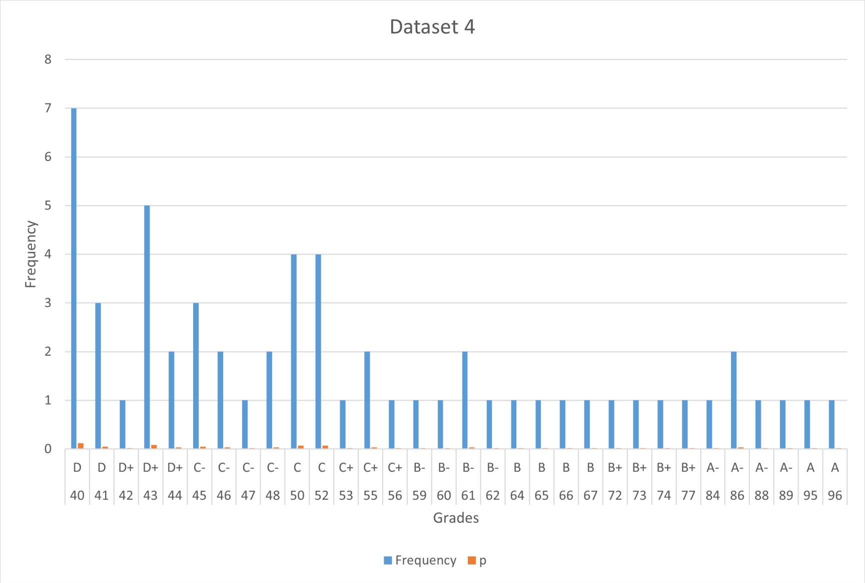


Figure 4: Numerical and Letter Grades of Dataset 1 (Fall 2021)

4.2 Model Results

To prepare the data in a proper way as the input of the model first, we have sorted the numerical grades increasingly. Second, deleted the data of the students who had dropped the course and their grade was N/A. Third, the vector of unique numerical grades was calculated by deleting the duplicates in order to calculate the probabilities (p) of each individual grade. This probability is the proportion of each number's frequency on the total number of grades (students). The model's inputs are the p vector and the number of bins which is set to be 10 as mentioned previously.

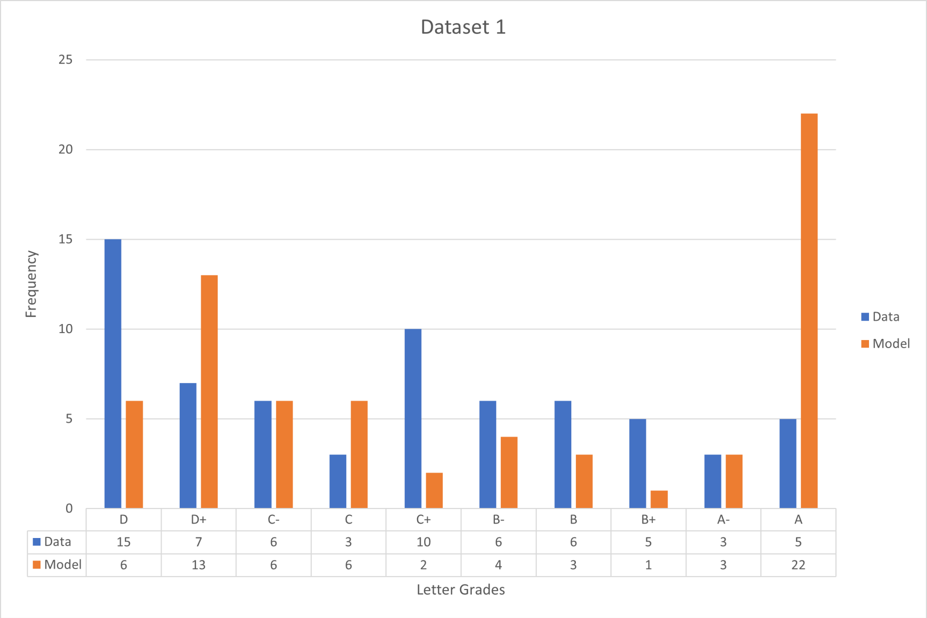
All experiments are implemented in Python 3.7 and optimization problems are solved by Gurobi 9.1.1 on an Intel(R) 3.7 GHz processor and 64 GB RAM workstation. The optimality tolerance is set to be 10^{-3} (0.1 percent) and the run time is limited to 30 minutes.

Table (1) shows the objective value and the run time of each experiment. As shown in the table, the model is just able to give the optimal solution for the dataset 4 in a reasonable amount of time and it reached the time limit for the rest.

Table 1: Computational Results

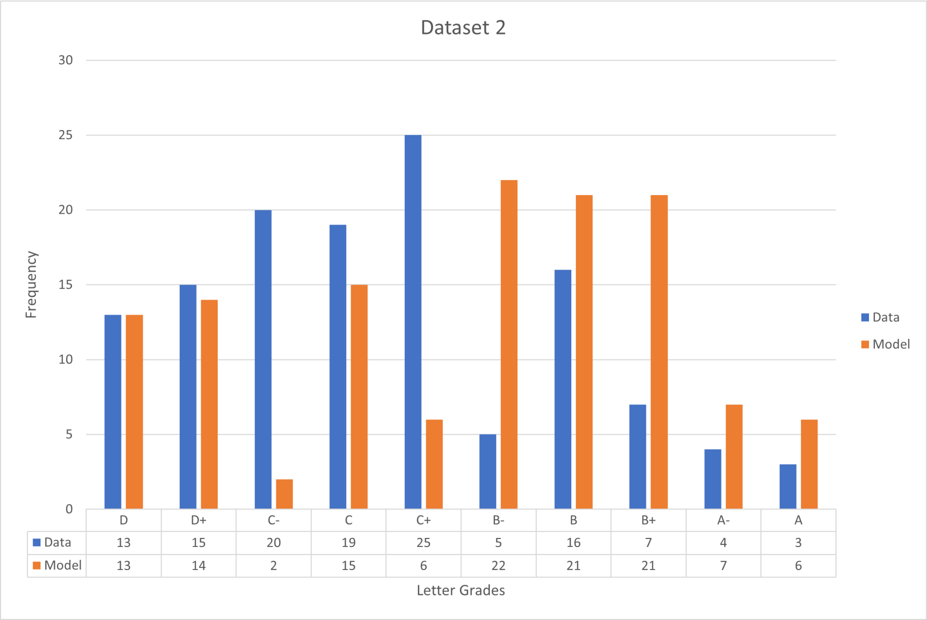
Dataset	Objective Value	Run Time
1	0.001226	1800.128
2	0.002509	1801.948
3	0.001184	1800.119
4	0.000853	516.3896

The frequencies of the letter grades in the datasets are compared with the models' results in figures (5)-(8). According to the letter grades associated with each numerical grade after the experiment by the model in table (2), for all the datasets model gives grades better or at least equal to those from the data that the instructors have given manually. In the other words, the model is more generous in assigning letter grades.



	D	D+	C-	C	C+	B-	B	B+	A-	A
🟦 Data	15	7	6	3	10	6	6	5	3	5
🟧 Model	6	13	6	6	2	4	3	1	3	22

Figure 5: Data Numerical and Letter Grades vs. Model Output - Dataset 1



	D	D+	C-	C	C+	B-	B	B+	A-	A
🟦 Data	13	15	20	19	25	5	16	7	4	3
🟧 Model	13	14	2	15	6	22	21	21	7	6

Figure 6: Data Numerical and Letter Grades vs. Model Output - Dataset 2

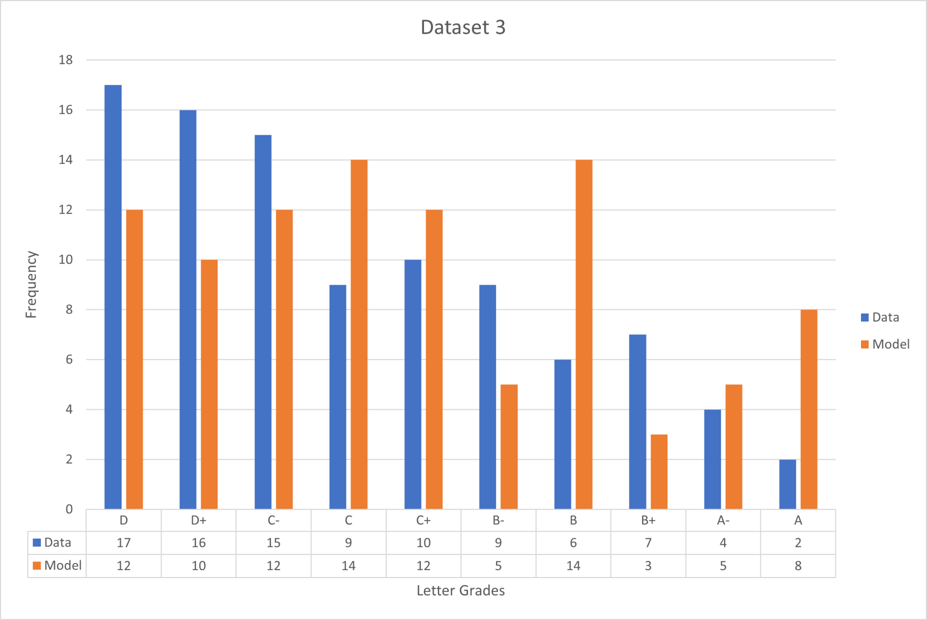


Figure 7: Data Numerical and Letter Grades vs. Model Output - Dataset 3

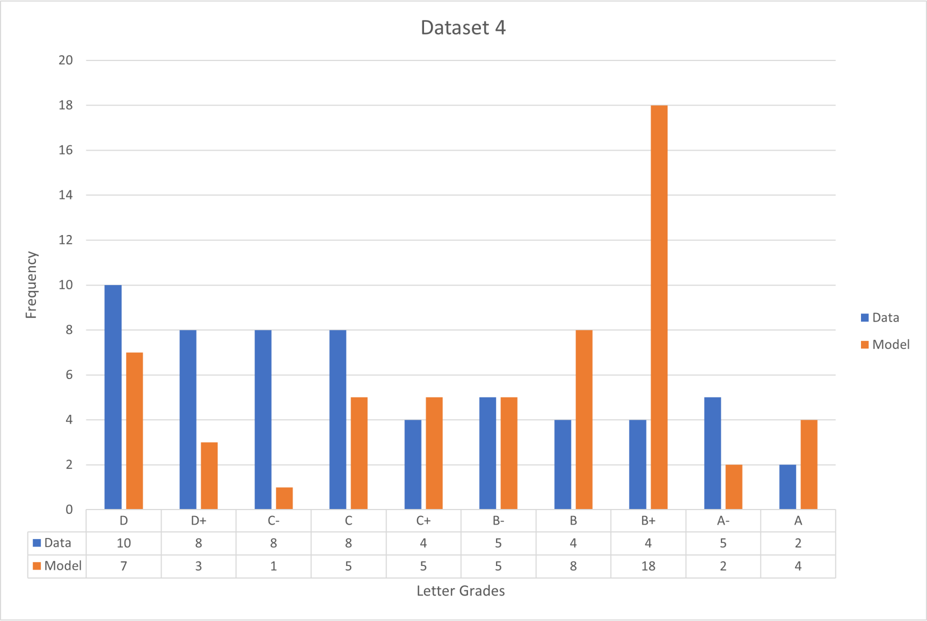


Figure 8: Data Numerical and Letter Grades vs. Model Output - Dataset 4

Table 2: Letter Grades of Data vs. Model Outputs

2[2]*Numerical	Dataset 1		Dataset 2		Dataset 3		Dataset 4	
	Data	Model	Data	Model	Data	Model	Data	Model
40	D	D	D	D	D	D	D	D
41	D	D	D+	D+	D	D+	D	D+
42	D	D+	D+	D+	D+	D+	D+	C-
43	D	D+	D+	D+	D+	C-	D+	C
44	-	-	D+	C-	D+	C-	D+	C+
45	D+	D+	C-	C-	C-	C-	C-	C+
46	D+	C-	C-	C	C-	C	C-	B-
47	D+	C-	C-	C	C-	C	C-	B-
48	C-	C-	C-	C	C-	C	C-	B-
49	C-	C-	C-	C+	C-	C	-	-
50	C-	C	C-	C+	C	C+	C	B
51	-	-	C	C+	C	C+	-	-
52	C	C	C	B-	C	C+	C	B
53	C+	C+	C	B-	C	C+	C+	B+
54	C+	C+	C	B-	C	C+	-	-
55	C+	B-	C+	B-	C+	C+	C+	B+
56	C+	B-	C+	B	C+	C+	C+	B+
57	C+	B	C+	B	C+	B-	-	-
58	C+	B+	C+	B	C+	B	-	-
59	B-	A-	C+	B	B-	B	B-	B+
60	B-	A	C+	B	B-	B	B-	B+
61	B-	A	B-	B	-	-	B-	B+
62	-	-	B-	B+	B-	B	B-	B+
63	B	A	B-	B+	B-	B	-	-
64	B	A	B	B+	B-	B	B	B+
65	-	-	B	B+	B-	B	B	B+
66	-	-	B	B+	-	-	B	B+
67	-	-	B	B+	B	B	B	B+
68	B	A	B	B+	-	-	-	-
69	B	A	B	B+	B	B+	-	-
71	-	-	B	B+	B	B+	-	-
72	B+	A	B	B+	B	B+	B+	B+
73	B+	A	B+	B+	-	-	B+	B+
74	B+	A	B+	A-	-	-	B+	B+
75	-	-	-	-	B+	A-	-	-
76	A-	A	A-	A-	B+	A	-	-
77	-	-	A-	A	-	-	B+	B+
78	A-	A	A-	A	B+	A	-	-
79	A-	A	A-	A	A-	A	-	-
80	-	-	-	-	A-	A	-	-
81	-	-	-	-	A-	A	-	-
84	-	-	A	A	-	-	A-	B+
86	A	A	A	A	-	-	A-	A-
88	-	-	-	-	-	-	A-	A
89	-	-	-	-	-	-	A-	A
91	-	-	-	-	A	A	-	-
95	-	-	-	-	A	A	A	A
96	-	-	-	-	-	-	A	A
98	A	A	A	A	-	-	-	-
100	A	A	-	-	-	-	-	-

5 Conclusion

In this work, the aim was to develop a mathematical model to construct the optimal histograms for assigning proper letter grades to numerical grades of students. The Wasserstein metric is utilized as a second-order cone representable distance measurement which has to be minimized subject to the satisfaction of some assumptions. In order to make the data ready and as the preprocessing step, the data which is the list of numerical grades should be cleaned and improper data must be substituted or removed. The next step is to remove the duplicates and sort the unique grades. By counting the number of repetitions of each unique grade, we have calculated the frequency of the grades in each dataset and made use of it to obtain the probability vector. This vector which is called p is the main input of the model. The other important input that we have to determine is the number of bins. Although the bin size and the number of bins could be two problems, in this research we set the number of bins as the number of letter grades which is 10 and let the model decide the bin size. The developed model is run for different datasets and the main key finding is that the model gives better grades compared to the instructors ones or equal to them. The complexity of the problem could be an issue since it is not able to solve 3 out of 4 datasets in the time limit of 30 minutes and it may be considered as a research area in the future.

References

[1] Hirotugu A. Information theory and an extension of the maximum likelihood principle. 1973.

[2] Taylor CC. Akaike's information criterion and the histogram. *Biometrika*, 1987.

[3] Scott DW. On optimal and data-based histograms. *Biometrika*, 66(3):605–610, 1979.

[4] Pearson K. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

[5] Knuth KH. Optimal data-based binning for histograms. *arXiv preprint physics*, 2006.

[6] Vaserstein LN. Markov processes over denumerable products of spaces, describing large systems of automata. *Problemy Peredachi Informatsii*, 5(3):64–72, 1969.

[7] Wand MP. Data-based choice of histogram bin width. *The American Statistician*, 1997.

[8] Kullback RA, Leibler. S. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.