Joseph Barbati
CS 378

Assignment 1
Data Exploration and Preprocessing

1a. Poker Hand Dataset

1b. Dataset of poker hands

- 1 table
- 10 attributes
    - 5 cards per hand
    - 2 attributes for each card (suit – ordinal, 1-4; rank – numerical, 1-13)
- 1,000,000 instances

1c. Classification is the perfect data mining application for this dataset. We are given 5 cards for each instance of a poker hand. Classifying it by what poker hand the cards give you (i.e. pair, flush, etc.) is exactly what this dataset needs.

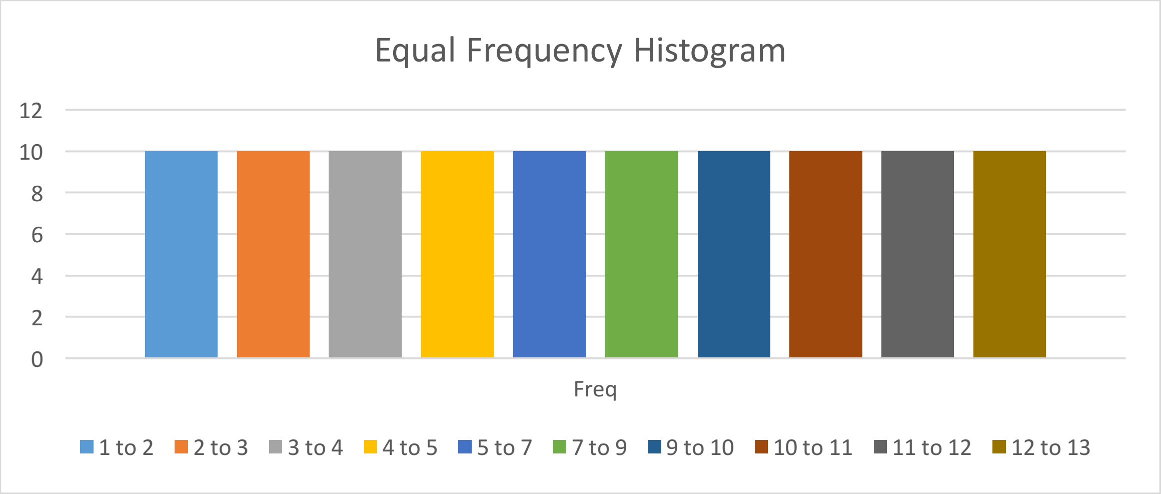
2a. Looking at C1 attribute (Rank of first card)

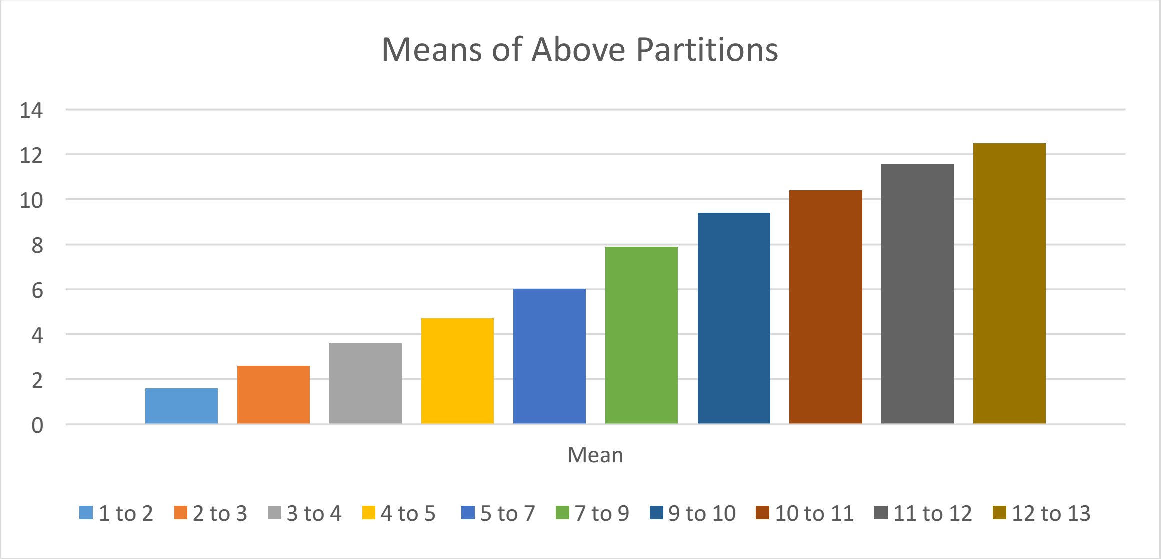
```
Mean: 6.997927
Median: 7
    1s: 77252
    2s: 76877
    3s: 76808
    4s: 77098
    5s: 76877
    6s: 77282
    7s: 76581
    8s: 76838
    9s: 76435
    10s: 76884
    11s: 77232
    12s: 76918
    13s: 76918
Mode: 6
Range: 12
Q1: 4
Q3: 10
Variance: 14.0128507154
Std Dev: 3.74337424197
```

2b. There don’t seem to be any quality issues with this data, barring any input errors. There is no missing data, and no unknown values.

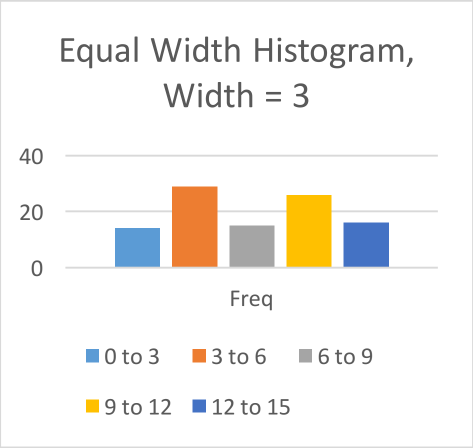
2c. As each number is important, representing a specific suit and rank, it is important not to mess with these values as they will skew the distribution of classes of poker hands. Therefore, I would say data smoothing is not necessary. Data reduction by data compression would be the best choice for preprocessing as 1M tuples is a lot to work with.

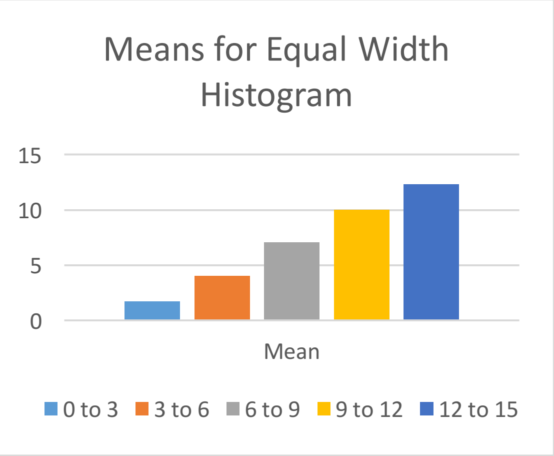
2c.



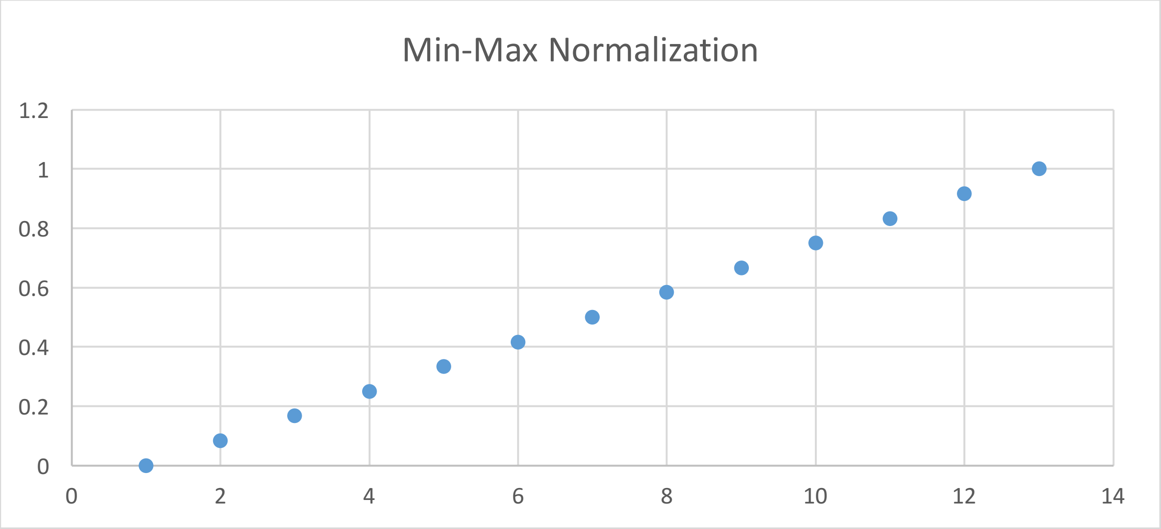


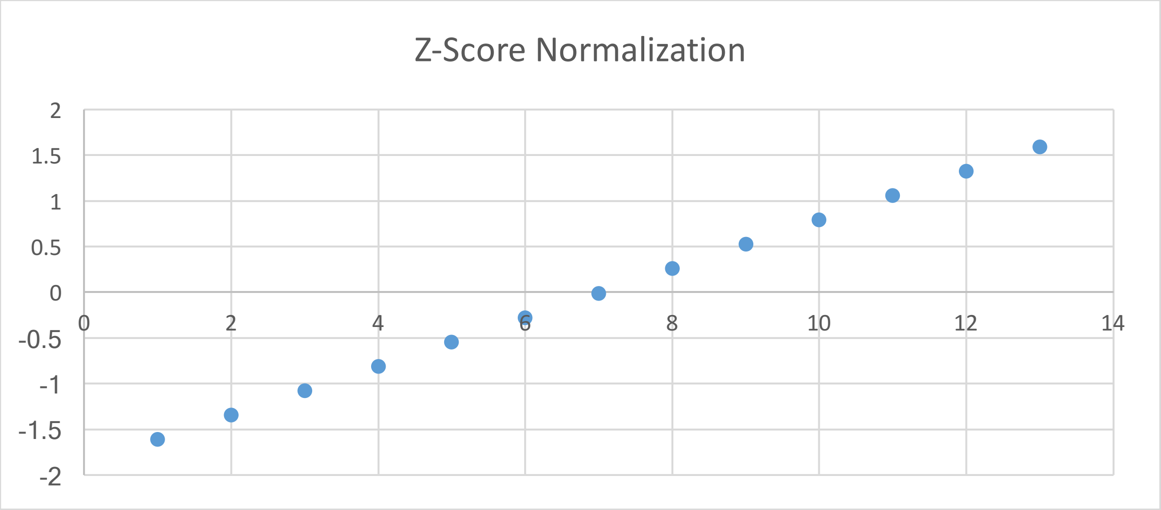
The above means would replace all values in the above equi-depth to perform data smoothing.





The above means would replace all values in the above equi-width histogram to perform data smoothing.





3b. C1 – C5 all have around the same distributions (similar to what is seen in 2a). S1 – S5 have virtually identical distributions, with an equal number of 1's, 2's, 3's, and 4's throughout S1 to S5. These observations are to be expected as each card is just as likely to be chosen the any other card at any given point when choosing from a 52 card deck.