HW1

Mustafa Sadiq (netID: ms3035)

1a)

population = U.S adults
sample = 1000 US adults

1b)

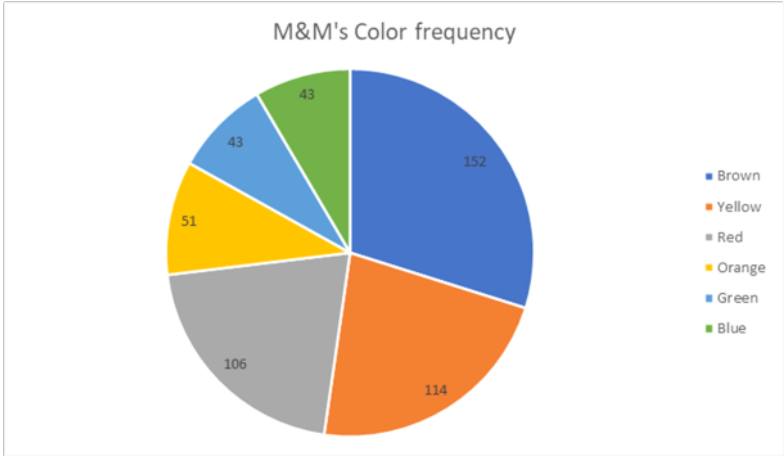
descriptive statistic since it is a fact and not an inference

2a)

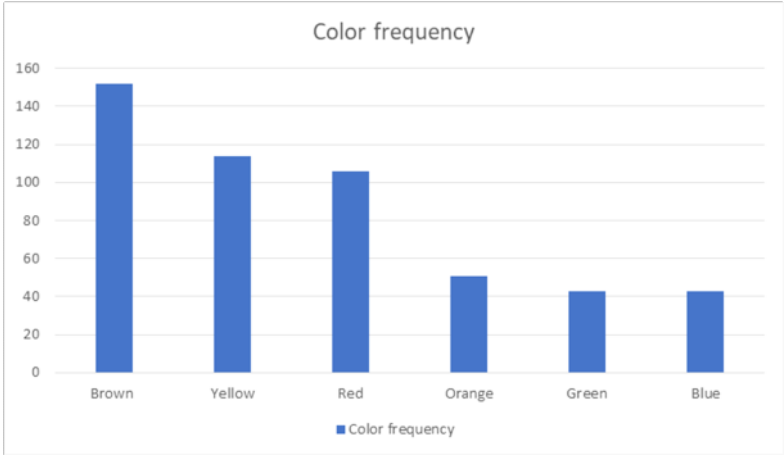
total = 152 + 114 + 106 + 51 + 43 + 43 = 509

Color	Frequency	Relative Frequency Distribution
Brown	152	152/509 = 0.29862
Yellow	114	114/509 = 0.22397
Red	106	106/509 = 0.20825
Orange	51	51/509 = 0.10020
Green	43	43/509 = 0.08448
Blue	43	43/509 = 0.08448

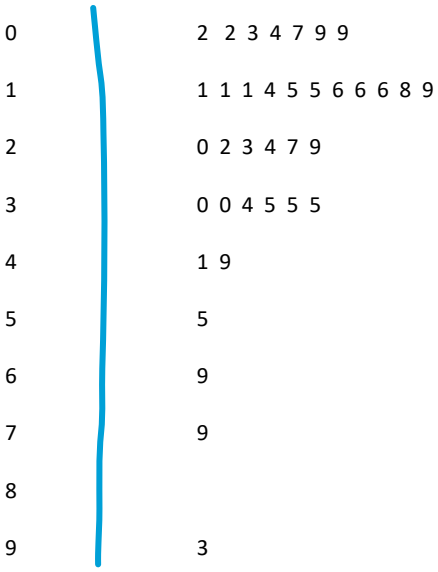
2b)



2c)



3a)



3b)

3c)

one line per stem as shape does not change much and adds many insignificant stems in the two line per stem

4a)

total = 83+64+46+48+523+35+34+265+2484+46+385+21+86+429+51+258+119 = 4977

mean = 4977/17 = 292.7647

21	34	35	46	46	48	51	64	83
86	119	258	265	385	429	523	2484	

17/2 = 8.5

median = 83

mode = 46

4b)

the median since we have a outlier calculated in our mean and mode is insignificant as no other repetitions.

5a)

mean dataset 1 = (1+5+1+8+2+8+2+9+5+9)/10 = **5**
mean dataset 2 = (1+9+1+9+1+9+1+9+1+9)/10 = **5**
mean dataset 3 = (5+5+5+5+5+5+5+5+5+5)/10 = **5**
mean dataset 4 = (2+4+4+4+4+4+4+4+10+4+10)/10 = **5**

5b)

They all have different variations.

5c)

Dataset 3 has least variation since all values are 5 which is the mean. Dataset 2 has the greatest variation since all values lie on extreme ends of the mean.

5d)

Range = highest value - lowest value

Dataset1 = 9-1 = **8**
Dataset2 = 9-1 = **8**
Dataset3 = 5-5 = **0**
Dataset4 = 10-2 = **8**

5e)

Sample standard deviation = $\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$

Dataset1 = $\sqrt{\frac{(1-5)^2+(5-5)^2+(1-5)^2+(8-5)^2+(2-5)^2+(8-5)^2+(2-5)^2+(9-5)^2+(5-5)^2+(9-5)^2}{10-1}} = \mathbf{3.333}$

Dataset2 = $\sqrt{\frac{(1-5)^2+(9-5)^2+(1-5)^2+(9-5)^2+(1-5)^2+(9-5)^2+(1-5)^2+(9-5)^2+(1-5)^2+(9-5)^2}{10-1}} = \mathbf{4.216}$

Dataset3 = $\sqrt{\frac{0}{10-1}} = \mathbf{0}$

Dataset4 = $\sqrt{\frac{(2-5)^2+(4-5)^2+(4-5)^2+(4-5)^2+(4-5)^2+(4-5)^2+(4-5)^2+(10-5)^2+(4-5)^2+(10-5)^2}{10-1}} = \mathbf{2.708}$

5f)

The **sample standard deviation** since we can't infer anything from range which is same for 3 datasets but data is very varied

5g)

Yes they are consistent. We can see that dataset 3 which is all 5s has 0 variance. Also dataset2 which has data all over the place has the highest variance.

6a)

Non-built-up as mean will be around 70 and values vary a lot.
Built-up has higher values with a mean around 90 and values don't vary a lot.

6b)

<u>Built-up</u>
Range = 103 - 69 = **34**

Total = 88+100+76+98+103+85+69 = 619
Mean = 619/7 = 88.4286

Standard deviation = $\sqrt{\frac{\left(88-\frac{619}{7}\right)^2+\left(100-\frac{619}{7}\right)^2+\left(76-\frac{619}{7}\right)^2+\left(98-\frac{619}{7}\right)^2+\left(103-\frac{619}{7}\right)^2+\left(85-\frac{619}{7}\right)^2+\left(69-\frac{619}{7}\right)^2}{7-1}} =$

12.791

Non-built-up
Range = 102 - 53 = **49**

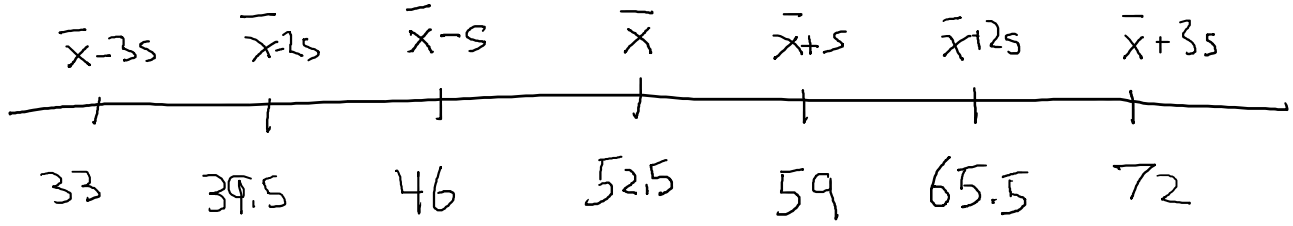
Total = 70+58+59+53+56+94+102 = 492
Mean = 492/7 = 70.2857

Standard deviation = $\sqrt{\frac{\left(70-\frac{492}{7}\right)^2+\left(58-\frac{492}{7}\right)^2+\left(59-\frac{492}{7}\right)^2+\left(53-\frac{492}{7}\right)^2+\left(56-\frac{492}{7}\right)^2+\left(94-\frac{492}{7}\right)^2+\left(102-\frac{492}{7}\right)^2}{7-1}} =$

19.788

Result is consistent with guess in part a. Non-built-up has a large range with data on extreme ends of the mean.

7a)



7b)

By Chebyshev's rule with k = 2, at least 75% of the people in the sample have early-onset dementia within two standard deviations to either side of the mean. Now, 75% of 21 is 15.75, and two standard deviations to either side of the mean is from 39.5 to 65.5, as we see from 7a).

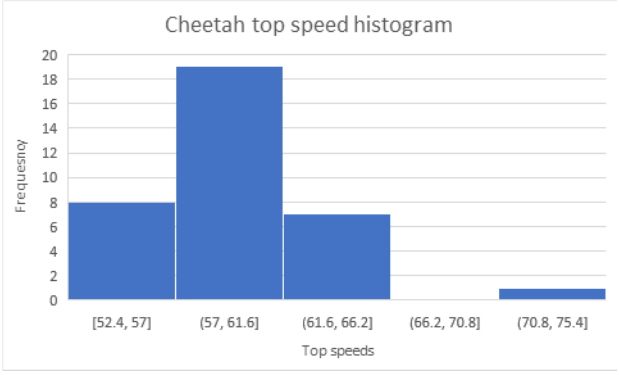
Interpretation: At least 16 (15.75 rounded up) of the 21 people in the sample have early-onset dementia between ages 39.5 and 65.5.

7c)

By Chebyshev's rule with k = 3, at least 89% of the people in the sample have early-onset dementia within three standard deviations to either side of the mean. Now, 89% of 21 is 18.69, and three standard deviations to either side of the mean is from 33 to 72, as we see from 7a).

Interpretation: At least 19 (18.69 rounded up) of the 21 people in the sample have early-onset dementia between ages 33 and 72.

8a)



According to the histogram, the data distribution is almost mound-shaped with insignificant

outliers. So we can apply the empirical rule to estimate the percentages of observations that lie within one, two and three standard deviations to either side of the mean

8b)

According to the empirical rule:

- Approximately 68% of the observations lie within one standard deviations to either side of the mean.
- Approximately 95& of the observations lie within two standard deviation to either side of the mean.
  Approximately 99.7% of the observations lie within three standard deviations to either side of the mean.

8c)

Sample mean = 59.53
Sample standard deviation = 4.27

<u>One standard deviation to either side of mean:</u>
59.53 - 4.27 = 55.26
59.53 + 4.27 = 63.8
Limits: 55.26 to 63.8
Observations:
55.4,55.5,55.9,56.5,57.3,57.5,57.6,57.8,57.8,58.1,58.3,58.7,59,59.2,59.6,59.7,59.8,60.1,60.2,60.6,60.7,60.9,61.3,61.6,62.3,62.6,63.4
Total observations = 27
Percentage = 27/35*100 = **77.1429**

<u>Two standard deviation to either side of mean:</u>
59.53 - 4.27 - 4.27 = 50.99
59.53 + 4.27 + 4.27 = 68.07
Limits = 50.99 to 68.07
Only one elements outside limits 75.3
Total observations = 35-1 = 34
Percentage = 34/35*100 = **97.1429**

<u>Three standard deviation to either side of mean:</u>
59.53 - 4.27 - 4.27 - 4.27 = 46.72
59.53 + 4.27 + 4.27 + 4.27 = 72.34
Limits = 46.72 to 72.34
Only one elements outside limits 75.3
Total observations = 35-1 = 34
Percentage = 34/35*100 = **97.1429**

8d)

A lot of observations lie within one standard deviationn to either side of the mean than estimated by the empirical rule. 68% estimated while 77.14% actual.

For observations within two standard deviation to either side of the mean, both are almost similar. 95% estimated while 97.14% actual.

There is no increase from two standard deviation to three standard deviation observations in the actual observations. It is less than the estimated value. 99.7% estimated while 97.14% actual.

9

- Which histograms are skewed to the left?
  - A, H
- Which histograms are approximately symmetric?
  - B, C, D, E, G
- For each of the four histograms A, B, C, and D, state whether the mean is likely to be larger than the median, smaller than the median, or approximately equal to the median.

- A: mean is likely less than the median
- B: mean is likely equal to the median
- C: mean is likely equal to the median
- D: mean is likely equal to the median

- Which of the distributions is likely to have the largest mean? The smallest mean?
  - Largest: C
  - Smallest: H

10.

Mean = 2.7
Standard deviation = 0.5

10a)

GPA = z * s.d + mean

z	gpa
2.0	2 * 0.5 + 2.7 = **3.7**
-1.0	-1 * 0.5 + 2.7 = **2.2**
0.5	0.5 * 0.5 + 2.7 = **2.95**
-2.5	-2.5 * 0.5 + 2.7 = **1.45**

10b)

-1.6 * 0.5 + 2.7 = **1.9**

10c)

According the empirical rule top 16% lies more than 1 standard deviation from the mean so z-score = 1
For top 2.5%, values lies more than 2 standard deviations from the mean so z-score = 2

For cum laude honors z-score = 1
Gpa = 1*0.5+2.7 = 3.2

For summa cum laude honors z-score = 2
Gpa = 2*0.5+2.7 = 3.7

To use the empirical rule we assumed that the student scores have a mound-shaped distribution and no significant outliers.