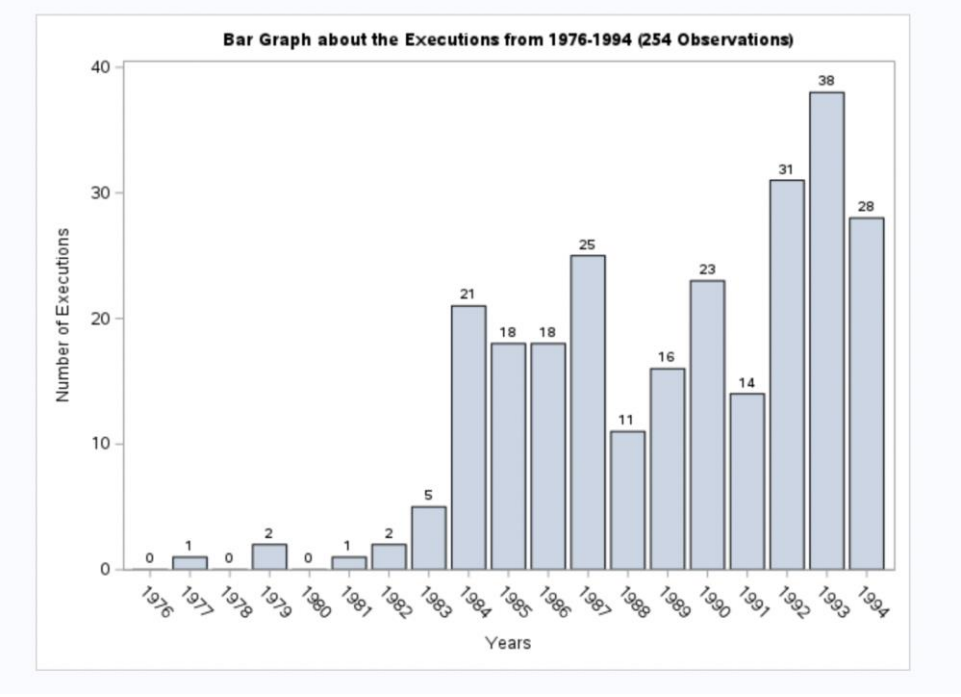
Homework 2

1.1: What types of graphs can be used to display nominal or ordinal observations? Discrete or continuous observations?

- Nominal observations are referring to categorical data that is often qualitative, such as marital status or color. Bar charts are good way to represent these observations since each bar can represent each category. Pie charts can also show the distinct categories found in the data.
- Similarly, ordinal observations are ordered categories. A bar chart is still useful, and it can be arranged in order.
- Discrete observations are quantitative, distinct values that can be shown as countable and whole numbers. Some variables like age can be discrete or continuous depending on if it is measured in whole numbers or decimals. To represent discrete observations, bar charts and pie charts can clearly show distinct numbers. A histogram can also be used if the data can be represented in intervals.
- Continuous observations have an infinite number of values between any range. They are often represented by decimals or other real numbers, so they offer precision for each value. Histograms and boxplots are often used to show these observations. Histograms put values in intervals or range and plots the number of observations for each interval. It is useful since it can also show the spread of data which is important for continuous data. A boxplot is useful since it can summarize the spread of data with the quartiles, median and outliers, while also showing the range of data across the graph.

1.2:



Statistics about the Executions from 1976-1994

The UNIVARIATE Procedure
Variable: exec (Number of Executions)

Moments			
N	19	Sum Weights	19
Mean	13.3684211	Sum Observations	254
Std Deviation	12.1207376	Variance	146.912281
Skewness	0.42531198	Kurtosis	-0.9742666
Uncorrected SS	6040	Corrected SS	2644.42105
Coeff Variation	90.666935	Std Error Mean	2.78068792

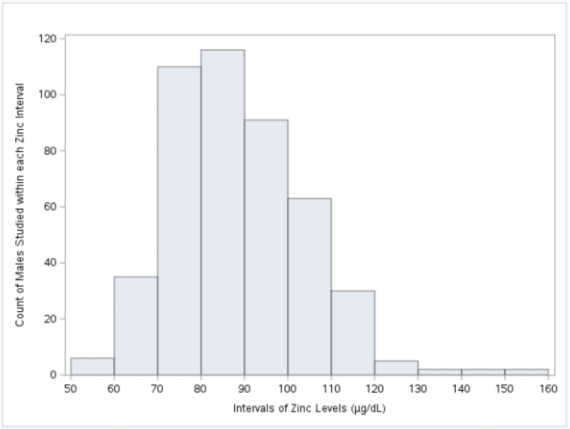
Basic Statistical Measures			
Location		Variability	
Mean	13.36842	Std Deviation	12.12074
Median	14.00000	Variance	146.91228
Mode	0.00000	Range	38.00000
		Interquartile Range	22.00000

The number of executions since 1976 has increased and varied greatly. A proc univariate procedure was done on the data and revealed that the mean is 13.37 while the standard deviation is 12.12. The standard deviation is close to the mean implying that there was a lot of deviation. Additionally, the most frequently found number is 0, but the IQR range is 22 which is a lot larger than the most found number in the dataset. Also, all numbers are positive.

1.3:

Histogram of Serum Zinc Levels for USA Males Ages 15-17 (n=462)

The UNIVARIATE Procedure



Summary of the Frequencies in Serum Zinc Levels (µg/dL)

Obs	zinc_int	males_sum	rel_freq
1	50-59	6	0.013
2	60-69	35	0.076
3	70-79	110	0.238
4	80-89	116	0.251
5	90-99	91	0.197
6	100-109	63	0.136
7	110-119	30	0.065
8	120-129	5	0.011
9	130-139	2	0.004
10	140-149	2	0.004
11	150-159	2	0.004

The relative frequencies show the mode or most frequent values. Based on the table on the right, the most common values were between 80-89 since this interval has the highest relative frequency in column 4 as 0.251. And, the data tends to center around the mode and resembles the bell-shaped curve of a normal distribution on the histogram. The data is not perfectly a normal distribution, but there is a clear center where a lot of data is located and the frequency of data tends to decrease as you move away from the center.

1.4: Under what conditions is use of the mean preferred? The median? The mode?

- Mean is preferred when the data resembles the normal distribution or bell-shaped, since the mean is the center of the data. Also, the mean can be greatly impacted by outliers, so if your data does not have a lot of outliers or if you would like to express the impact of outliers, the mean is a good measure.
- Median is preferred if the data contains outliers since it is impacted less by outliers or extreme values. Or, if the data does not resemble the normal distribution or is skewed, the median can be a better representation than mean.
- Mode is preferred for categorical or discrete data since it can show the most frequently found category or interval. It can be good for histograms or frequencies.

1.5:

Descriptive Statistics of Serum Zinc Levels Dataset for USA Males Ages 15-17

The MEANS Procedure

Analysis Variable : zinc					
N	Mean	Std Dev	Median	Lower Quartile	Upper Quartile
462	87.9372294	16.0046888	86.0000000	76.0000000	98.0000000

More Statistics of Serum Zinc Levels Dataset for USA Males Ages 15-17

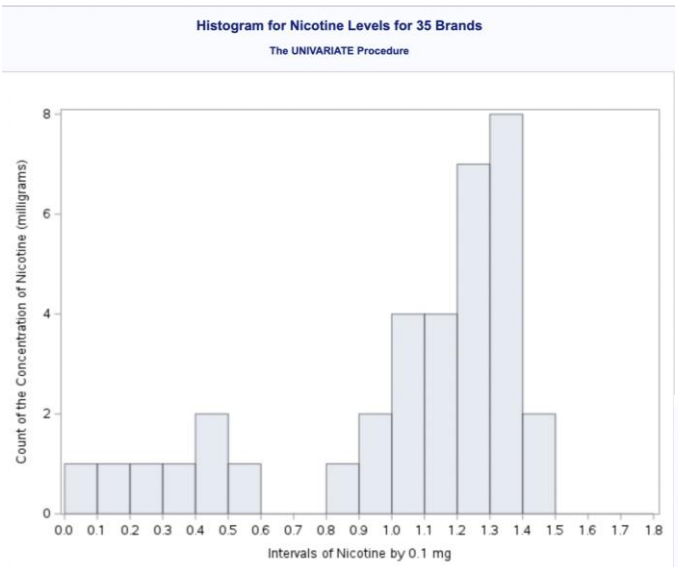
The UNIVARIATE Procedure
Variable: zinc

Moments			
N	462	Sum Weights	462
Mean	87.9372294	Sum Observations	40627
Std Deviation	16.0046888	Variance	256.150064
Skewness	0.62315144	Kurtosis	0.91287614
Uncorrected SS	3690711	Corrected SS	118085.18
Coeff Variation	18.2001286	Std Error Mean	0.74460551

Basic Statistical Measures			
Location		Variability	
Mean	87.93723	Std Deviation	16.00469
Median	86.00000	Variance	256.15006
Mode	75.00000	Range	103.00000
		Interquartile Range	22.00000

The histogram resembles a normal distribution, and the empirical rules say that one would expect about 95% of the values to be within 2 standard deviations of the mean, and 99.7% of the values to be within 3 standard deviations of the mean.

1.6:



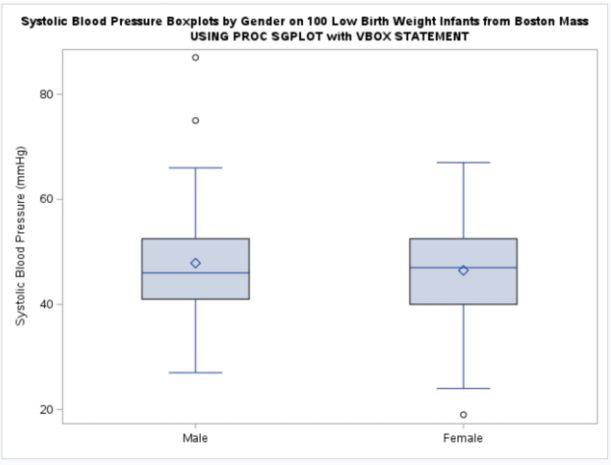
Some Descriptive Statistics for Nicotine Levels

The MEANS Procedure

Analysis Variable : nicotine			
N	Mean	Median	Std Dev
35	0.9908571	1.1000000	0.3883326

The mean is 0.99 and median is 1.1. These values are very close to each other, so both of these values can work to represent the dataset. It slightly resembles a normal distribution, but a larger sample than 35 brands could show the bell-shaped curve better. However, the graph also seems more skewed to the right with lots of outliers on the leftmost side. To account for this skewness, the mean is the best value for central tendency because it accounts for the outliers or skewness.

1.7:



Mean, Median, Count and Standard Deviation of Systolic Blood Pressure (mmHg) for Each Gender

The MEANS Procedure

Gender=0

Analysis Variable : sbp Systolic Blood Pressure (mmHg)			
N	Mean	Median	Std Dev
56	46.4642857	47.0000000	11.1452628

Gender=1

Analysis Variable : sbp Systolic Blood Pressure (mmHg)			
N	Mean	Median	Std Dev
44	47.8636364	46.0000000	11.8057749

Both male and female boxplots have a normal distribution since the median and mean of the systolic blood pressure for both gender are very close. The mean systolic blood pressure for the males is 47.8 mmHg and median is 46 mmHg. The mean systolic blood pressure for the females is 46.4 mmHg and median is 47 mmHg. The mean can show the presence of outliers and median is another measure of central tendency, so if the values are similar, then the data follows bell-shaped curve. The females boxplot has one outlier at about 19 mmHg, and males boxplot has two outliers in 70's and 80's.

1.8: Here is a table of the Summary Statistics:

Table: Summary Statistics of the Female Low Birth Weight Infants (n=100)

	Statistics	*Female (n=56)*	*Male (n=44)*
Systolic Blood Pressure (mmHg)	Mean	46.5	47.9
	St Dev	11.1	11.8
Gestational Age (weeks)	Mean	28.9	28.9
	St Dev	2.1	2.8
APGAR Score (5 minutes)	Mean	6.1	6.4
	St Dev	2.5	2.3
Toxemia			
No	N (%)	45 (80.4%)	34 (77.3%)
Yes		11 (19.6%)	10 (22.7%)
Germinal Matrix Hemorrhage			
No	N (%)	45 (80.4%)	40 (90.9%)
Yes		11 (19.6%)	4 (9.1%)