Homework 1.2
Part2

Chaonan Shi
1901412;

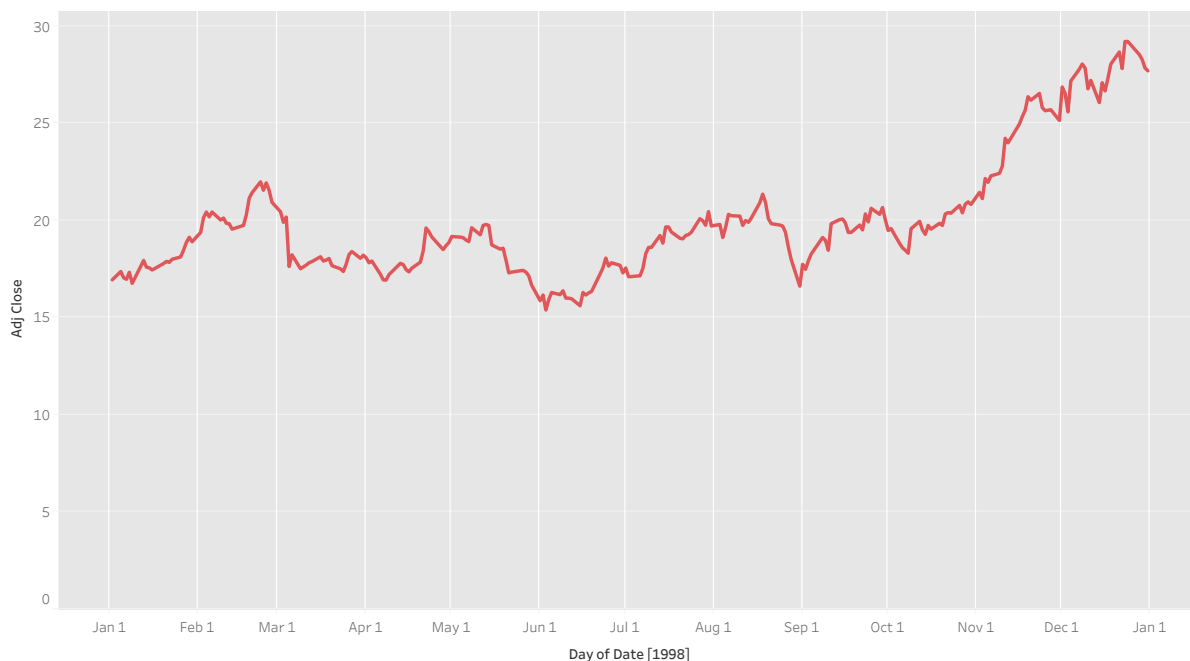***Problem 2:***
A:
For Tableau:

Figure 2.1
Adj.Close vs. the date

Day of Date [1998]

The trend of sum of Adj Close for Date Day.

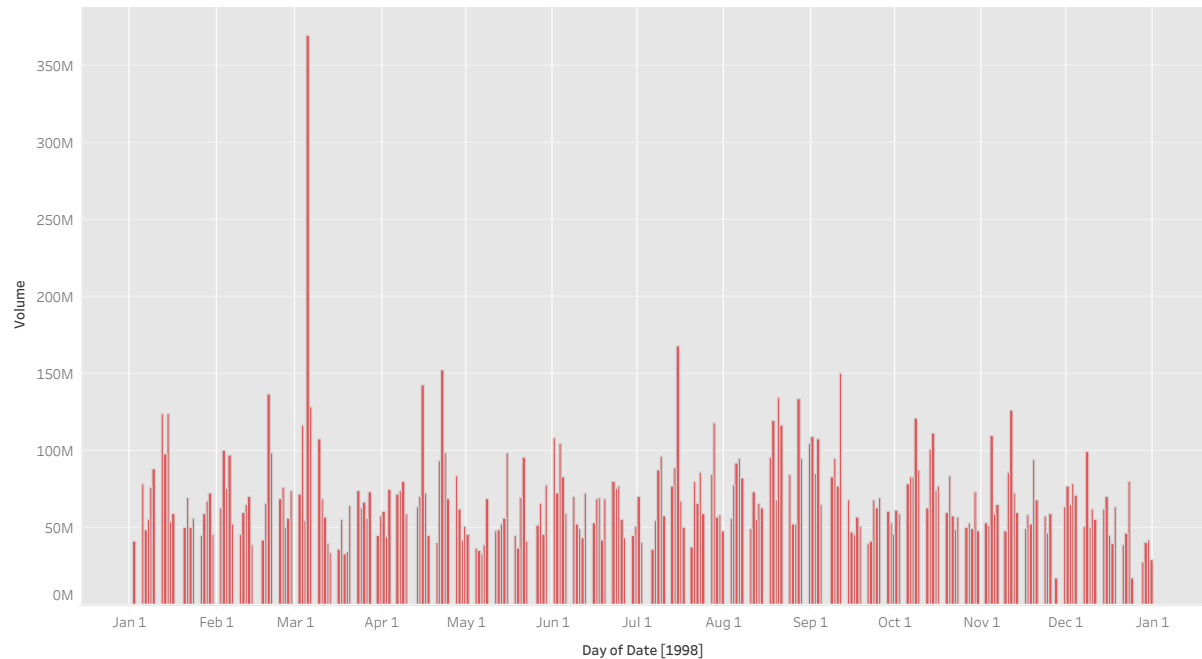For tableau, I choose color of the background as grey, and trending line as red.
In the summary, the Title is adj.close VS date from Jan 1998 – Dec 1998;
The Horizontal axe represents the Day of Data [1998] as exact data, and the Vertical axe represents the #Adj.close;
In this case, we can summarize that the trend of Adj.Close (adjusted closing price) significantly raised from Oct 1998 to Jan 1999;

B:
For Tableau:

Figure 2.2
Volume vs the date

The plot of sum of Volume for Date Day.

For tableau, I choose color of the background as grey, and trending bar as red;
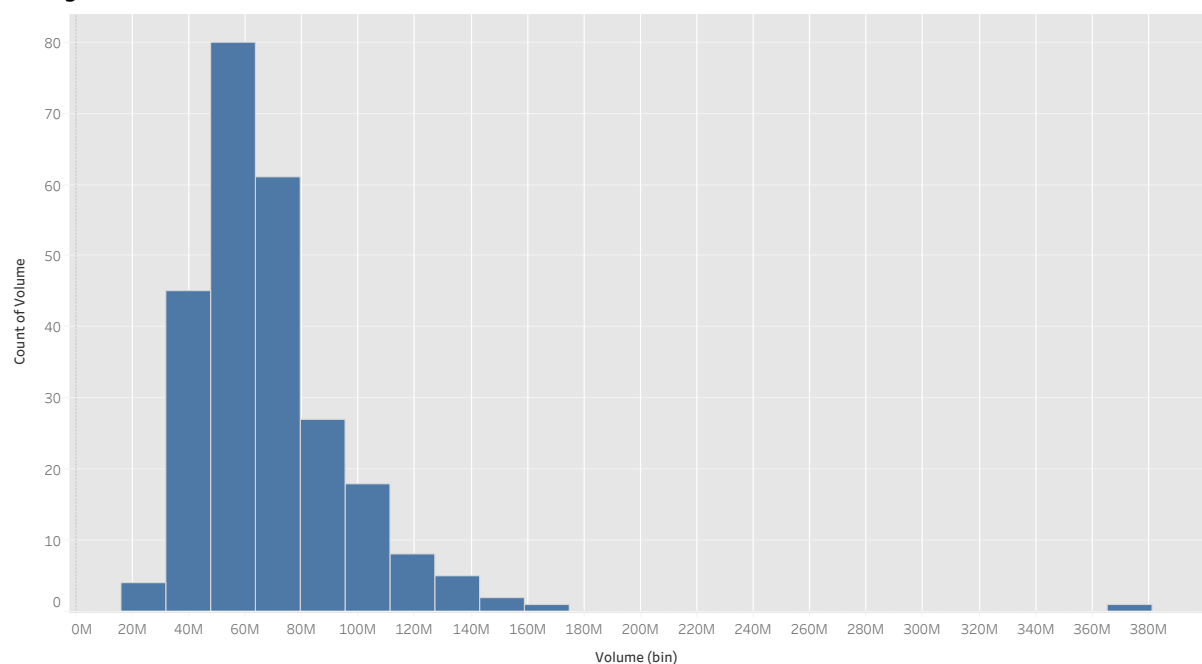In the summary, the Title is Volume VS date from Jan 1998 – Dec 1998;

The Horizontal axe represents the Day of Data [1998] as exact data, and the Vertical axe represents the Volume (shares traded);

In this case, we can summarize that the trend of Volume (shares traded) significantly higher than other date at Mar 1 at 1999;

C:
For Tableau:

Figure 2.3

Histogram for Volume



The trend of count of Volume for Volume (bin).

For tableau, I choose color of the background as grey, and histogram as blue;
In the summary, the Title is Histogram for Volume at 1998;
The Horizontal axe represents the Volume (shares traded) as exact data, and the Vertical
axe represents the Count of Volume;

In this case, we can summarize that the shape of volume are not normally distributed,
tending to be **skew to the right** with existing **outliers** at 380M;

D:
For Tableau:

Figure 2.4

Scatterplot for Volume VS daily price range

Volume vs. Range. Color shows details about Range (bin).
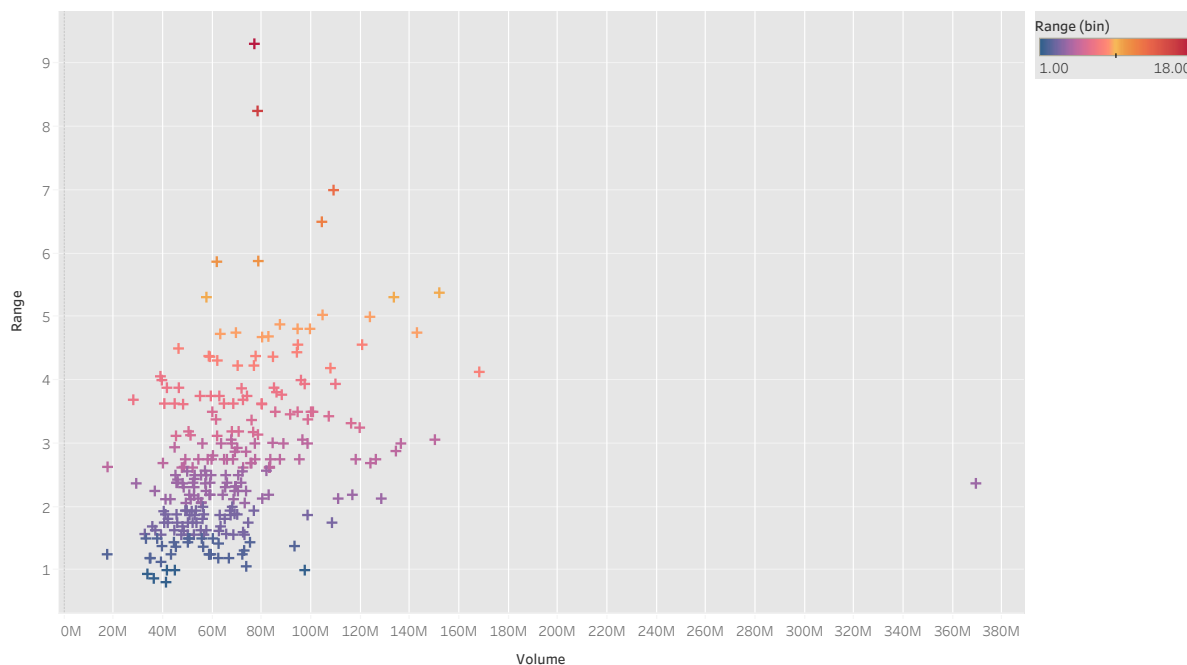
For tableau, I choose color of the background as grey, and color of observations of volume based on the Volume (from blue to red), and shape of observations from circle to to the "plus".
In the summary, the Title is Scatterplot for Volume VS daily price range;

The Horizontal axe represents the Volume (shares traded) as exact data, and the Vertical

axe represents the range (high - low);

In this case, we can summarize that the shape correlation of Volume and Range is not linear, but tend to be positive, which means that the higher volume may have bigger range. Moreover, there are some existing **outliers** at 80M and 380M;

Problem 3:

a:

For Tableau:

Figure 3.1

Error		×

[Response] – [True Value]

▶

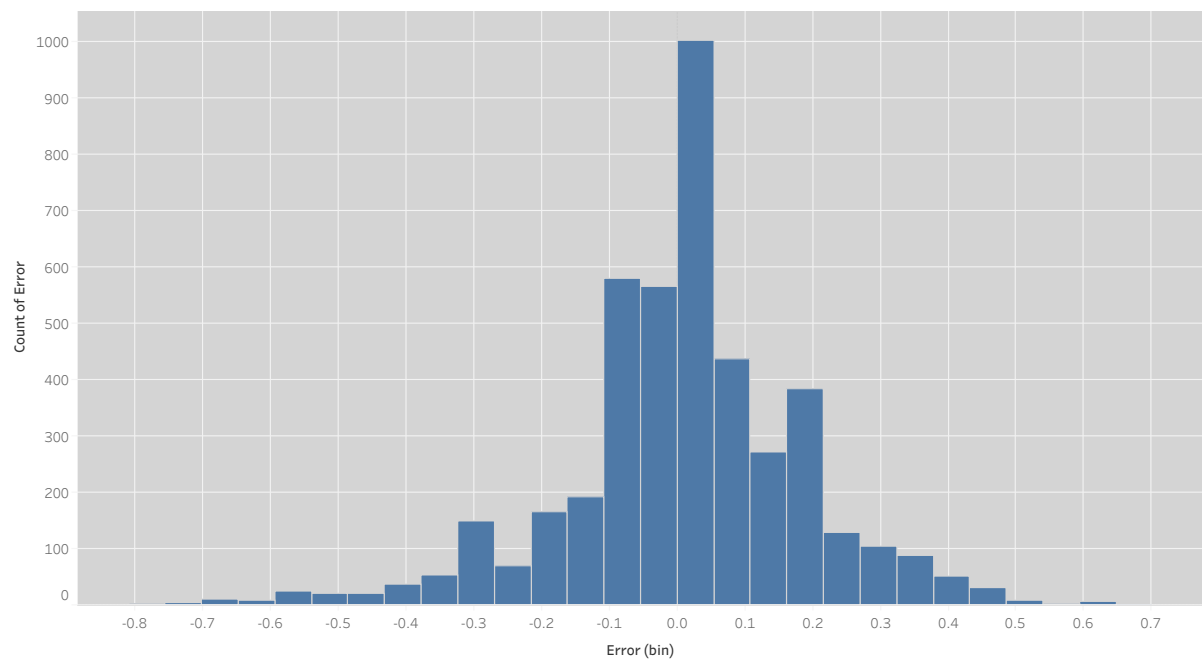The calculation is valid.   2 Dependencies ▾   Apply   OK

Create a new column that contains the amount of error: Error = Response – TrueValue;

b:

For Tableau:

Figure 3.2

Histogram for Error

The trend of count of Error for Error (bin).

For tableau, I choose color of the background as grey, and histogram as blue;

Moreover, I **slightly narrowed down** the width of bar for histogram for better display;
In the summary, the Title is Histogram for Error;

The Horizontal axe represents the Error (Response – TrueValue) and the Vertical axe represents the Count of Error;

In this case, we can summarize that the shape of error is not perfectly normal distribution, we may **need to transform** the Error for getting better result;

c:
For Tableau:

Figure 3.3
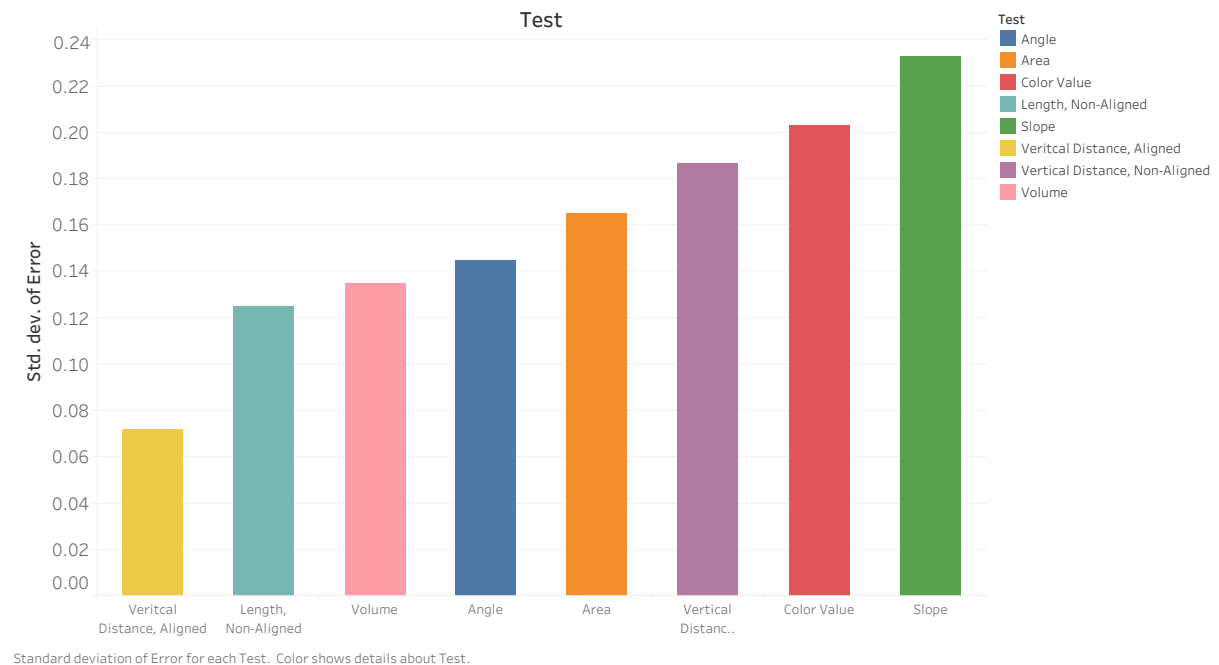Median *Error* vs *Test*
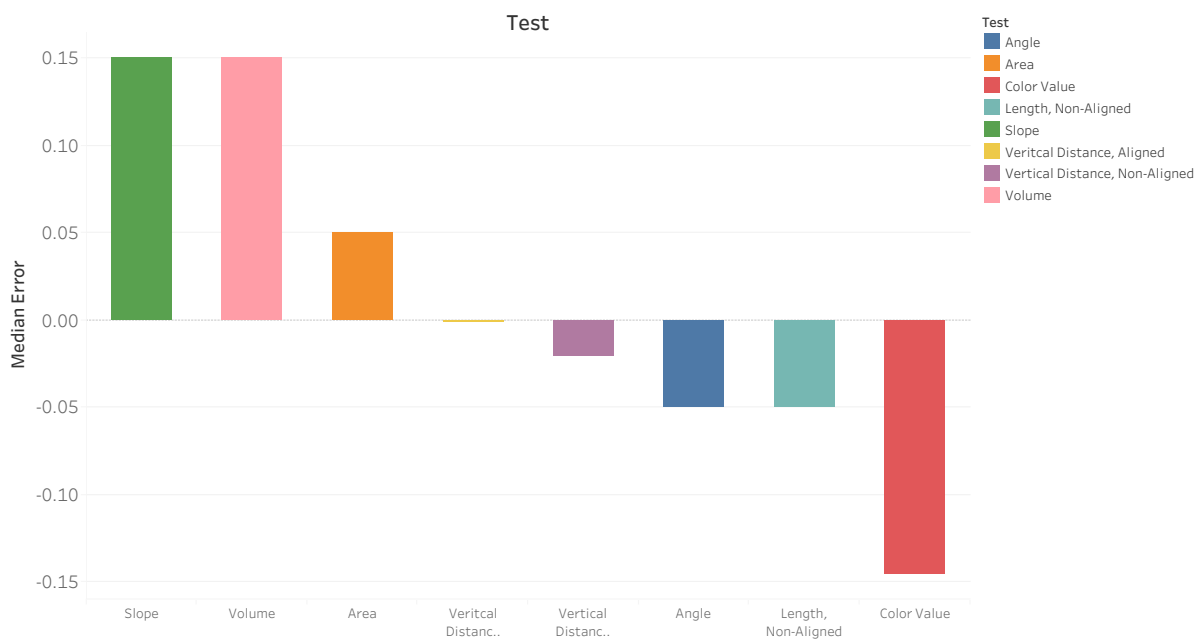
Median of Error for each Test. Color shows details about Test.

For tableau, I choose color of the background as white, and colors of test for bar;
Moreover, I **enlarged** the width of bar for better display; and set up the Error from default(SUM) to Median;
In the summary, the Title is median Error vs. Test;

The Horizontal axe represents the Test and the Vertical axe represents the Count of Error;

In this case, we can summarize that Slope, Volume, and Color Value have higher Median Error compare with others.

d:
For Tableau:

Figure 3.4
Standard Deviation vs *Test*

Standard deviation of Error for each Test. Color shows details about Test.
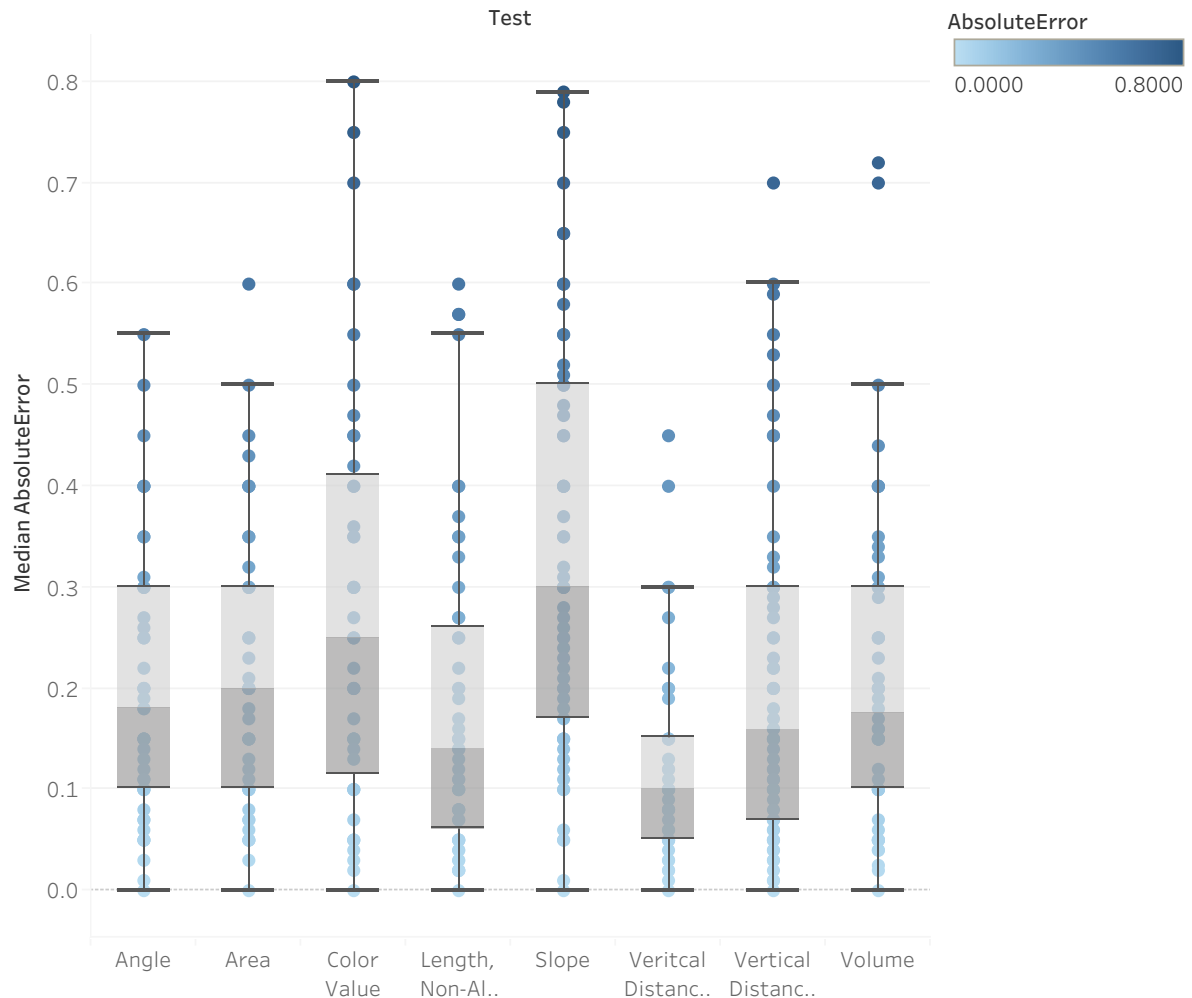
For tableau, I choose color of the background as white, and colors of test for bar;
Moreover, I **sorted** the order of bar for better display, from low to high, and set up the Error from default(SUM) to Standard Deviation;
In the summary, the Title is Standard Deviation VS Test;

The Horizontal axe represents the Test and the Vertical axe represents the Std of Error;

In this case, we can summarize that Slope, Vertical Distance Aligned, and Color Value have higher Std Error compare with others. And the Vertical Distance Aligned has lower Std Error, on the contrast.

e:

For Tableau:

Figure 3.5

 boxplot for AbsoluteError by Test.

Median of AbsoluteError for each Test. Color shows details about AbsoluteError.

For tableau, I choose color of the background as white, and colors of AbsoluteError for bar; Moreover, I **adjust the Test as dimension and AbsoluteError as Median** for bar to get better display

In the summary, the Title is boxplot for AbsoluteError by Test.;

The Horizontal axe represents the Test and the Vertical axe represents the Median

AbsoluteError;

In this case, we can summarize that Slope, Vertical Distance Aligned, and Color Value have higher Std Error compare with others. And the Vertical Distance Aligned has lower Std Error, on the contrast.
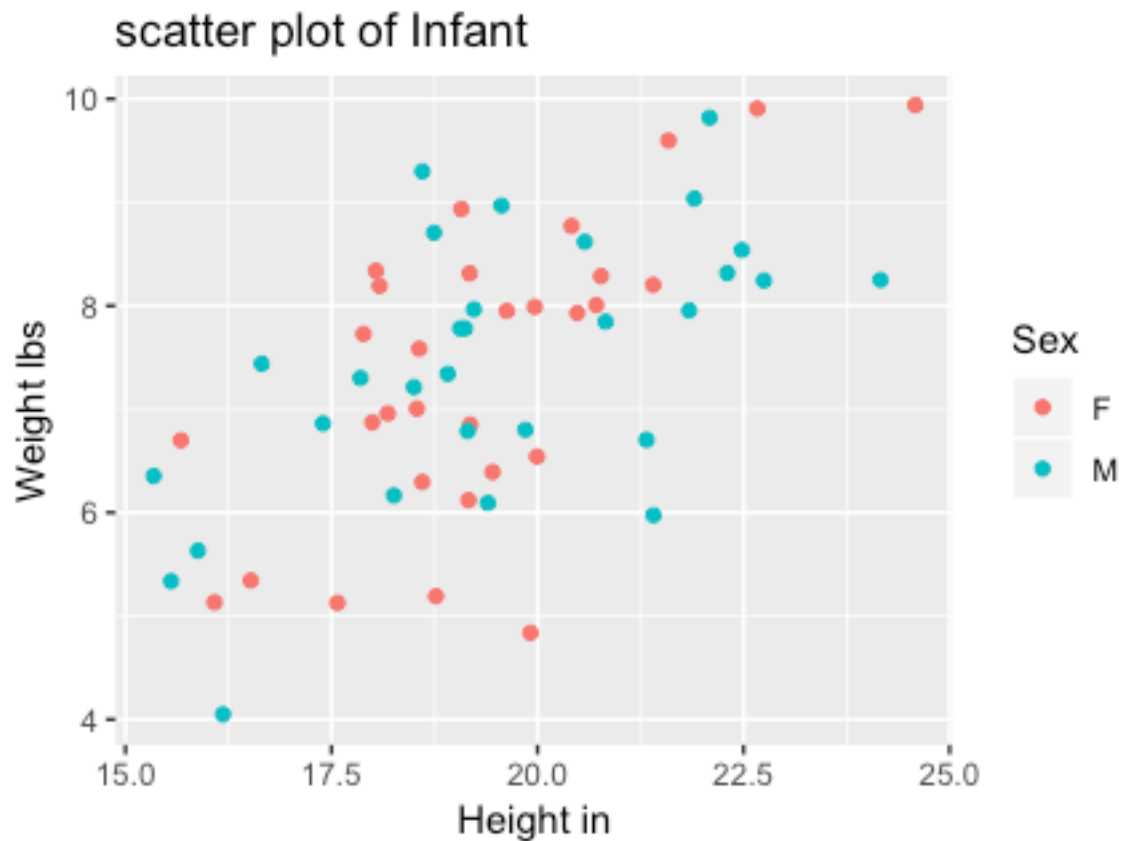
Problem 4:

a:

For R:

Figure 4.1

scatter plot of Infant

For R, I choose color of the background as grey, and colors of Height in and Weight lbs based on the Sex;
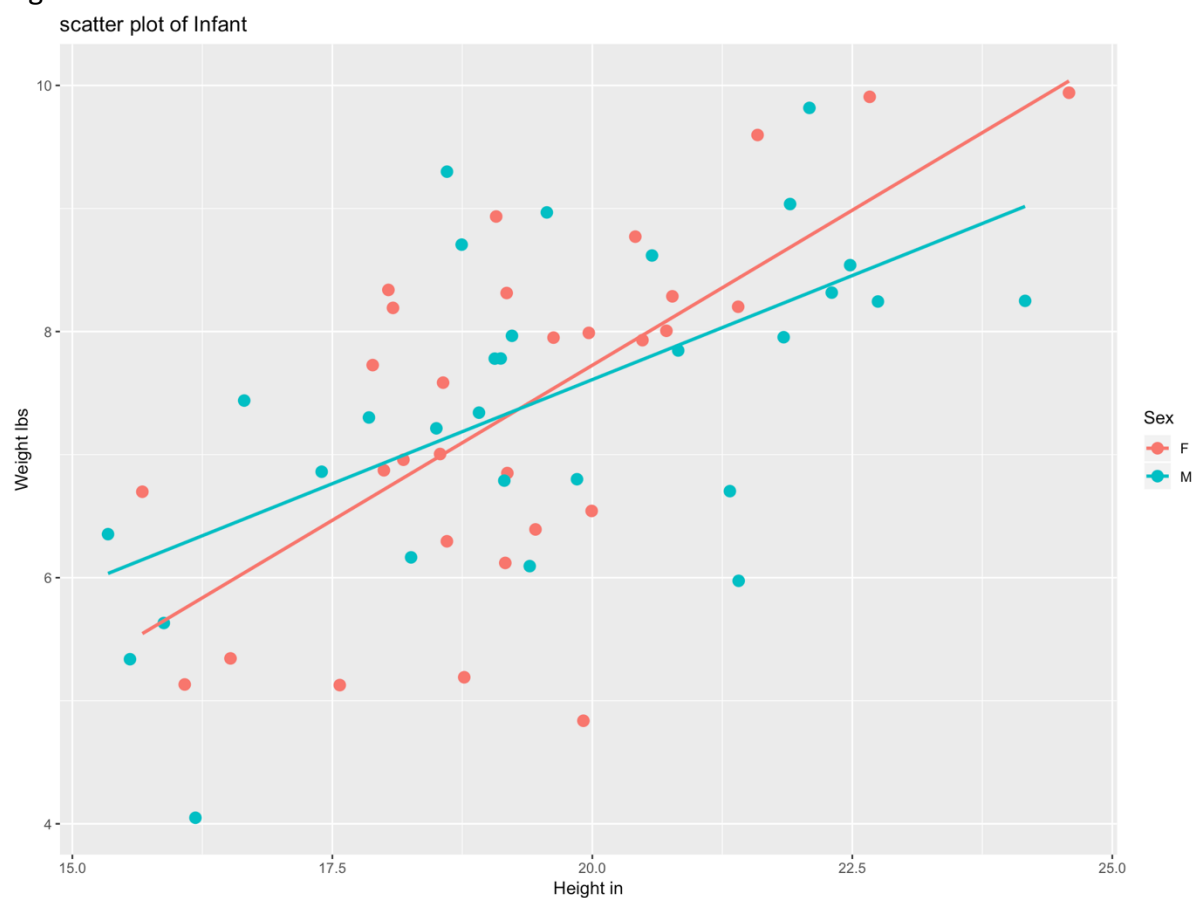
In the summary, the Title is boxplot for scatter plot of infant.;

The Horizontal axe represents the Height in and the Vertical axe represents the Weight lbs;

In this case, we can summarize that the shape correlation of Height in and Weight lbs is not linear, but tend to be positive, which means that the Height in may have bigger Weight lbs.

b:
For R:

Figure 4.2

scatter plot of Infant

For R, I choose color of the background as grey, and colors of Height in and Weight lbs based on the Sex;
In the summary, the Title is boxplot for scatter plot of infant.;
The Horizontal axe represents the Height in and the Vertical axe represents the Weight lbs;

In this case, I adjusted the size = 3, and added separate trend lines for the two populations on the graph plotted.


c:
Based on the above graphs, we can summarize that both of Height in and Weight lbs have positive slope, and the slope of F (female) is steeper than the M (male).

I enlarge the observation points to make them more readable. Moreover, I tried to summarize the graph with different shapes, so triangle represents Male, and circle represents the Female.
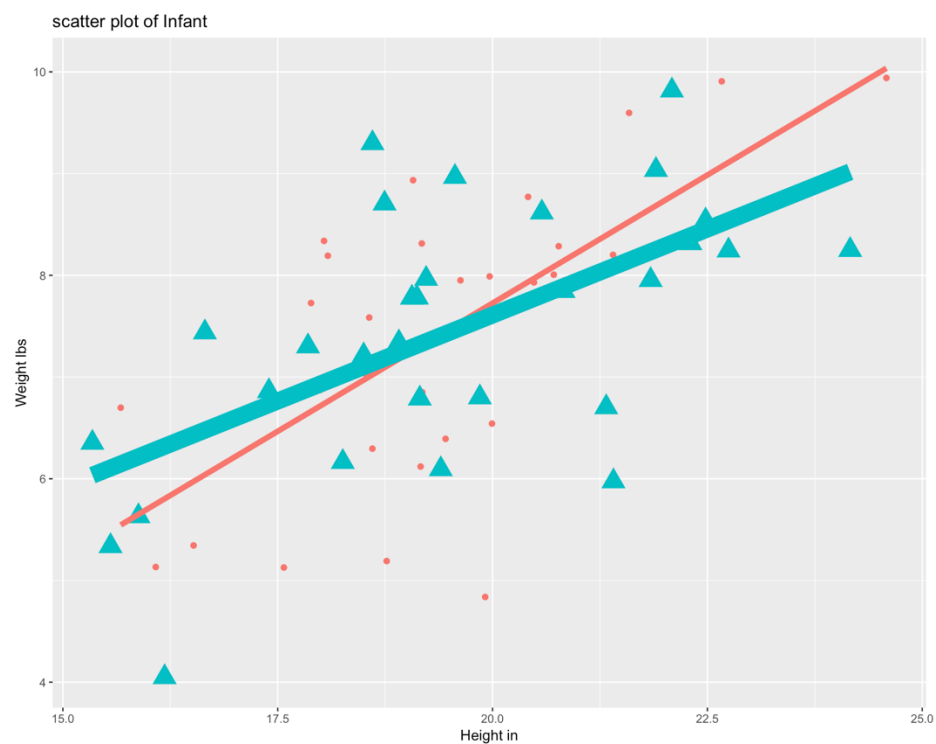
For R:

Figure 4.3

scatter plot of Infant
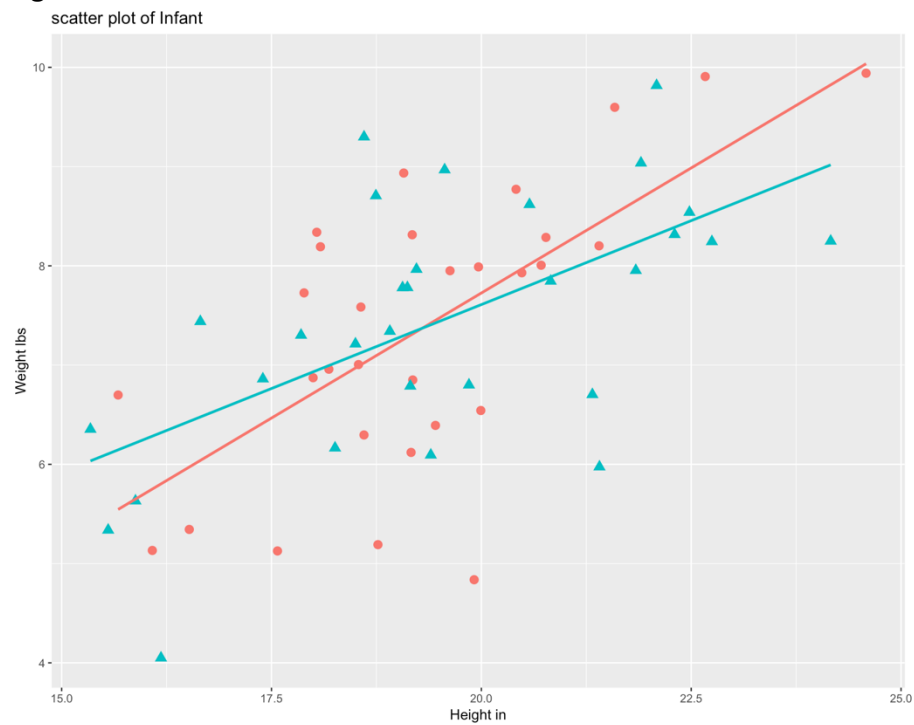
In addition, I tried to separated the lines by Sex, however, the result make graph unreadable, for example, the line of Male became larger overlapping some of observations in the graph (see Figure 4.4). So, personally, **I suggest to Figure 4.3 as my result**.
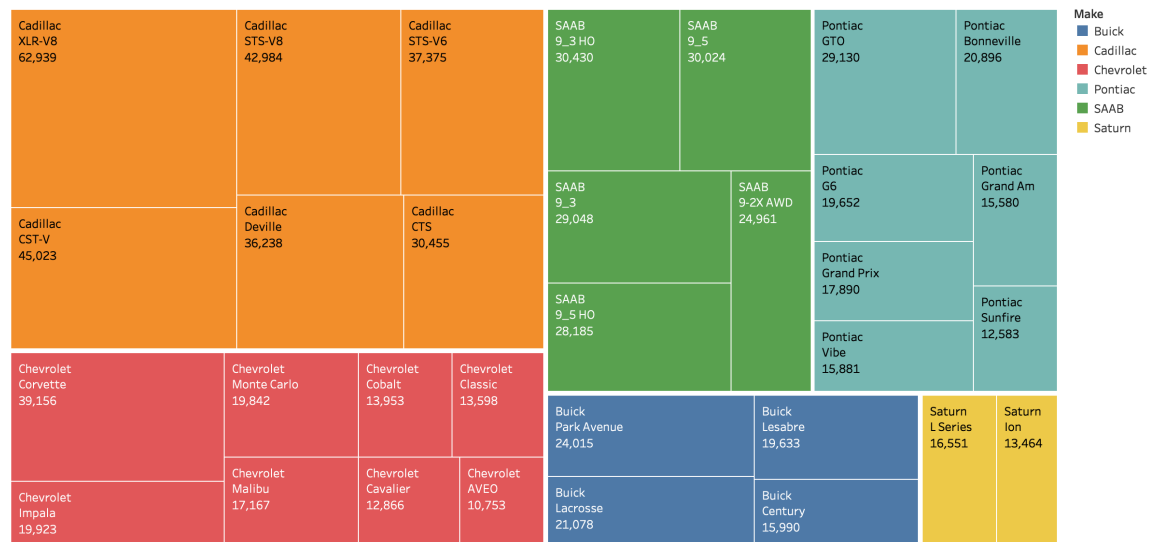
Figure 4.4

scatter plot of Infant

Problem 5:

a:

For Tableau:

Figure 5.1

Treemap for Price, Model, and make

Make, Model and average of Price. Color shows details about Make. Size shows average of Price. The marks are labeled by Make, Model and average of Price.

I clicked Show Me on the toolbar, then selected the treemap chart type. The title is bold displaying in center of figure. In order to show all information including Make, Mode, and price in my figure, I selected three of them as label and adjusted their size. The Make is also selected as color that we can see the same Make is in same color. What's more, the mean is used instead of using sum.

In this case, we can summarize below information:

Make	Model	Avg. Price
Cadillac	XLR-V8	62,938.736572

Which selling with **highest price**. And:

Make	Model	Avg. Price
Saturn	Ion	13,464.3838302

Which selling with **lowest price;**

b:
For Tableau:

Figure 5.2

packed bubble for Price, Model, and make

Make and Model. Color shows details about Make. Size shows average of Price. The marks are labeled by Make and Model.

For packed bubble graph, we got same information with previous question. However, since my screen is too small, some information been hidden.

c:
Based on above TWO different graph, we can summarize that the treemap provide more information for us. For example, treemap provide information with clear subdivision and hierarchy. Even more, it sorted the data from large to small (the top left is most expensive one, and bottom right is least expensive model);

For packed bubble, intuitively, this graph provide might be more clear to people as the first glance: the big bubble represents expensive one, and the small represents cheap one. However, as showing above, it might hide information since the hardware limited. For my computer, some of information cannot be displayed because the screen limitation. Moreover, from my point of view, the packed bubble might be too cute, and treemap could be more professional.
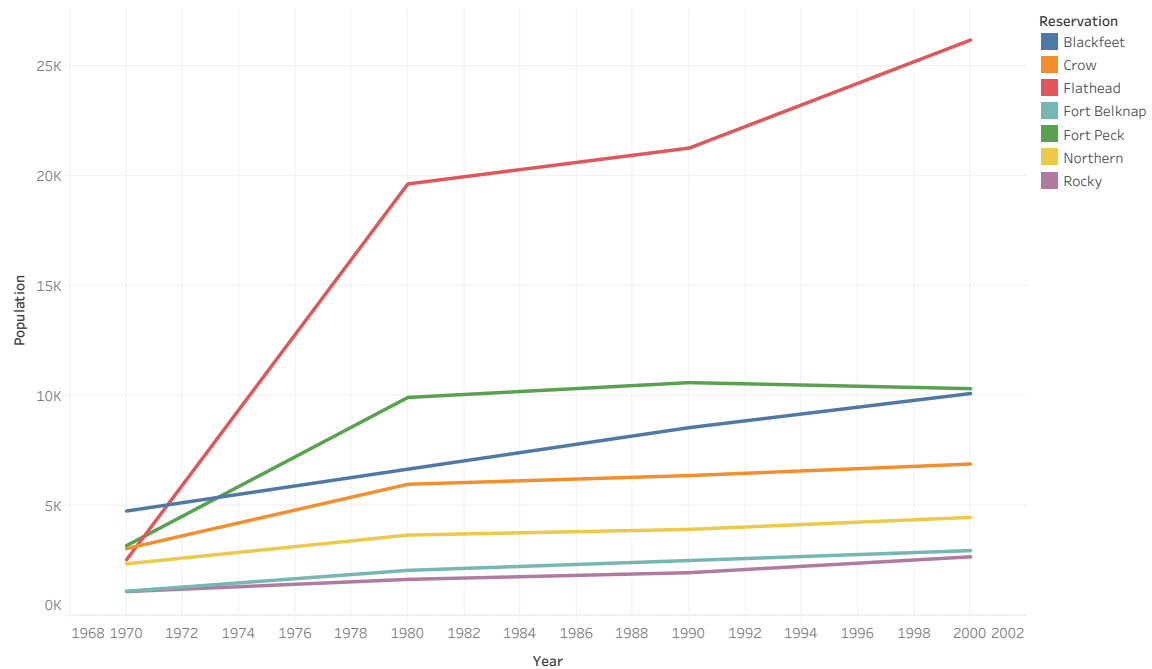
Problem 6:

a:

For Tableau:

Figure 5.1

population growth over the years for reservation

The trend of sum of Population for Year. Color shows details about Reservation.

For tableau, I choose color of the background as white, and colors of Reservation for lines;

In the summary, the Title is population growth over the years for reservation.;
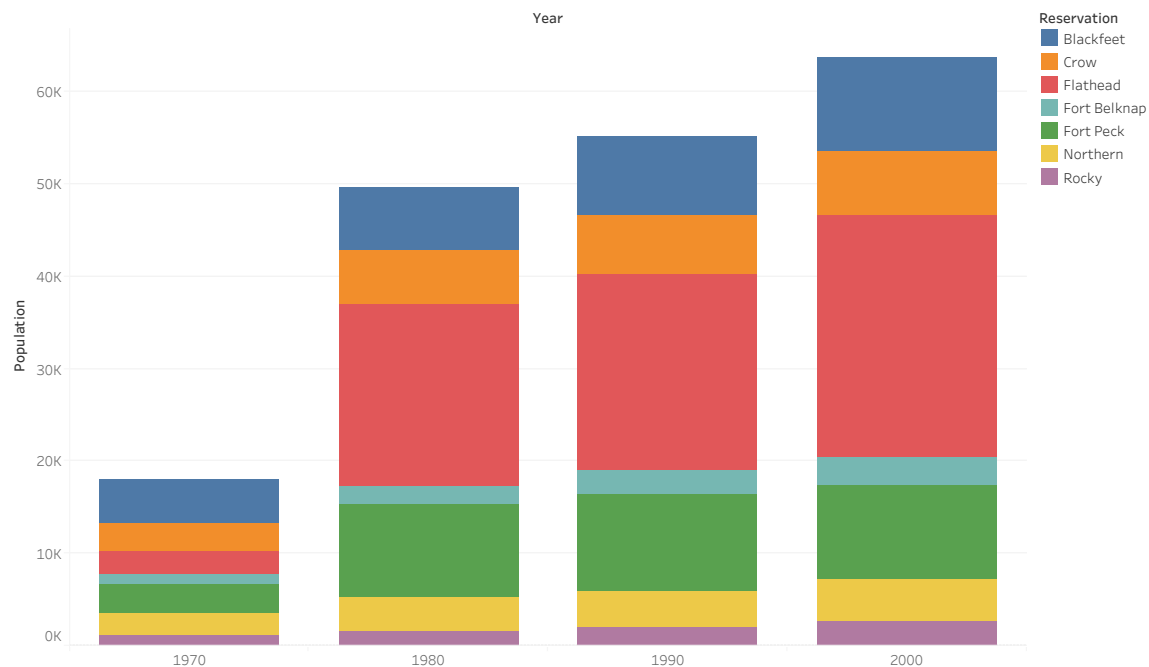
The Horizontal axe represents the Years from 1968 – 2002 and the Vertical axe represents the population growth;

In this case, we can summarize that the population of reservation of Flathread increased significantly than others; and blackfeet start with highest intercept with roughly 5k. On the contrast, the reservation of rocky with lowest growth and intercept.

b:
For Tableau:

Figure 5.2
total reservation population for each year

Sum of Population for each Year. Color shows details about Reservation. The data is filtered on Year, which keeps all values.

For tableau, I choose color of the background as white, and colors of Reservation for bars;

In the summary, the Title is Total reservation population for each year, for better review, I discrete the graph of year so we can get information only by decades;
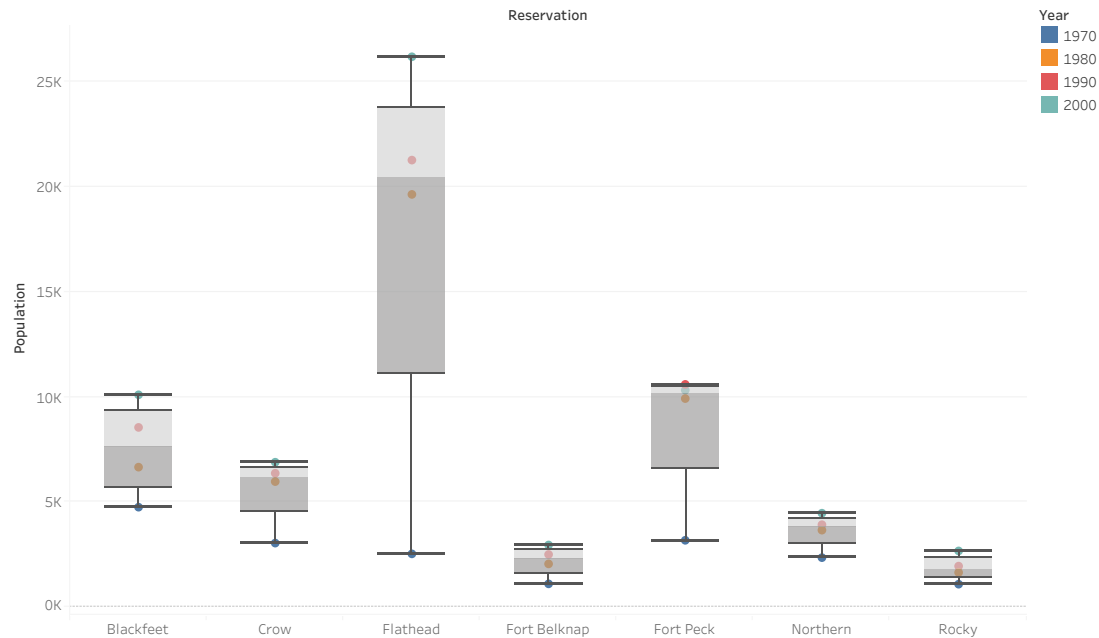The Horizontal axe represents the Years from 1970 – 2000 and the Vertical axe represents the population growth;

The results are same as pervious question, but we also focus at the total reservation population for each decades.

c:
For Tableau:

Figure 5.3

Boxplot

Population for each Reservation. Color shows details about Year. Details are shown for Reservation.

For tableau, I choose color of the background as white, and colors of year for boxplot;

In the summary, the Title is population distribution vs. years.
The Horizontal axe represents the categories of reservation and the Vertical axe represents the population growth;

The results are same as pervious question, but we also focus on the individual information statistically. For example, each box shows the Whisker, Hinge, and median. The points inside of the box represent the different year for the growth.

Problem 7:

Graph issues:
From my point of view, the graph obviously messed up the dimensions. For example, people would get trouble with depth of earthquake with probability, based on what we learned in the class, I would like to use simple X axe to represent the time of year, not staircase.

Second, labels of description are not clear. In this case, earthquake happened frequently between 1850 and 1906, however, the description of each earthquake was hidden by each other. In other works, it is hard to look through the detail about each earthquake. In this case, I would like to use the TreeMap of bar chart to describe the information about each earthquake.

Then the indicator of increasing quake probability not match the graph. People would consider the X-axe of staircase as increasing quake probability based on the description:
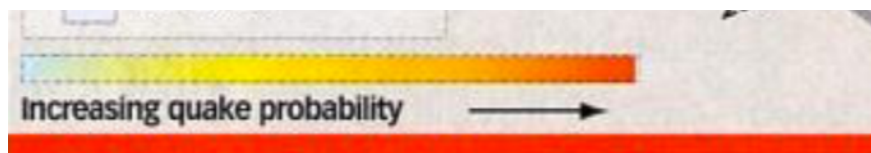
However, the actual X-axe of staircase was measured by depth of earthquake:
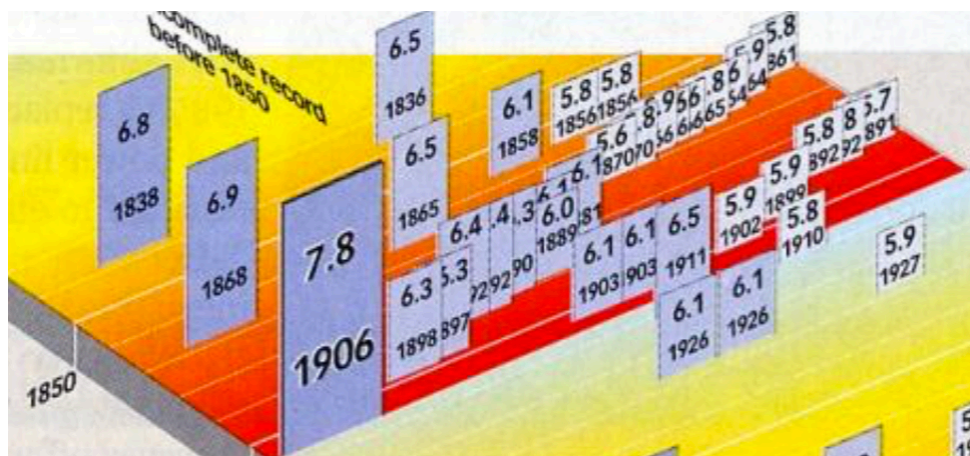
Again, as mentioned in the class, fancy would not always be better. Too many dimensions would not help people to read the graph efficiently. Personally, I would like to use the TWO dimensions graph and color basis to describe the depth, time, and frequency.

Finally, the levels of earthquake are not intuitive. Generally the 7 level's earth quake would has much more power of destruction than the level 6. So intuitively, this graph would not provide the sense of destruction by each level of earthquake.