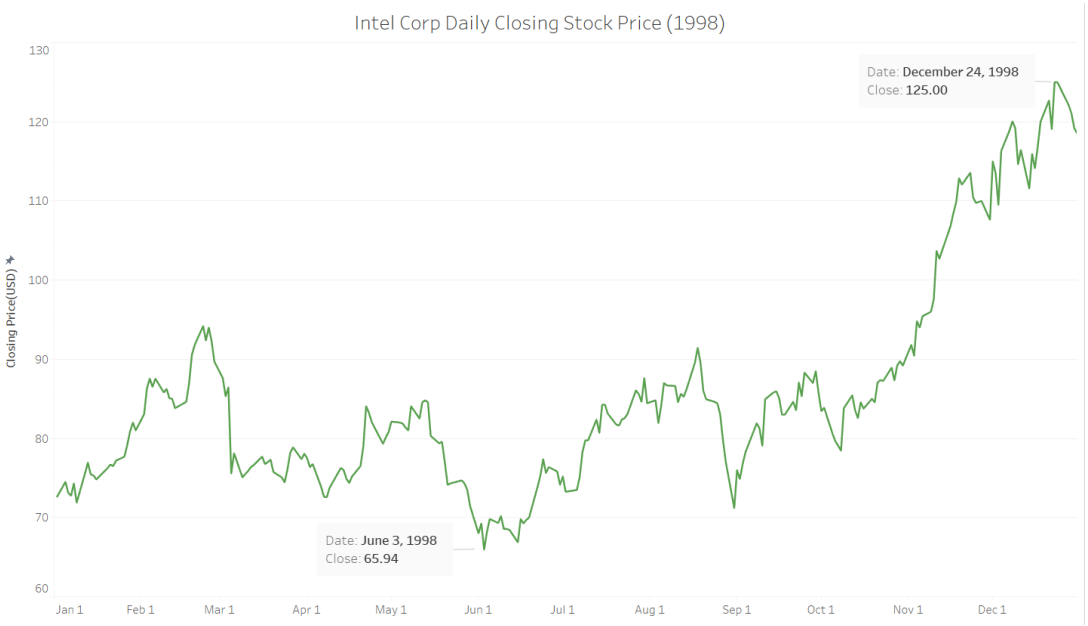
Elizabeth Nieto
HW 1

Problem 1: Trading Day

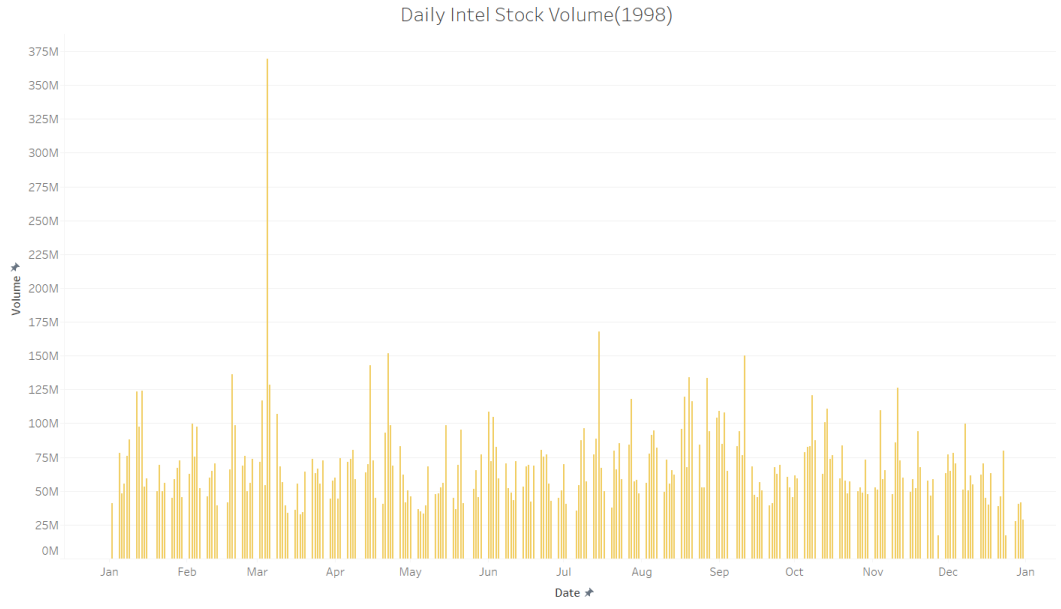
a. *Graph the closing price vs. the date with an ordinary line graph. If you use Tableau, you need to right-click on the Date and choose Exact Date from the dropdown menu so that it uses the full date with "day".*

Steps: For this graph I moved the date to columns and set it to exact date and then moved the closing price to rows. I then identified the highest and lowest points for the year and annotated them.



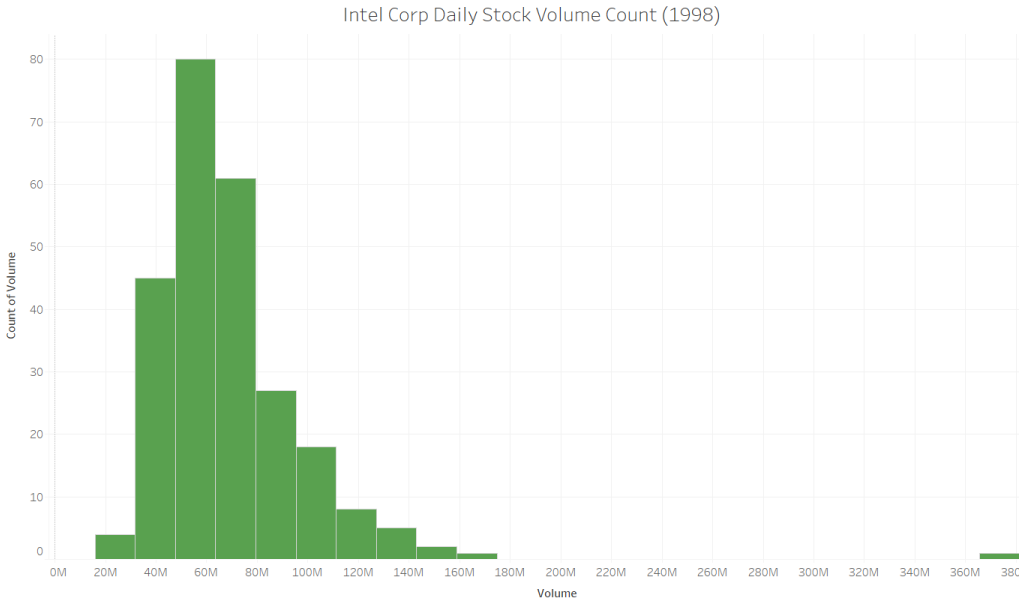
b. *Graph the Volume vs. the exact Date as in the last part with a bar graph.*

Steps: To create this graph date was moved to the columns and the volume sum to the rows. Bar width was adjusted to remove overlap among graphed dates. The color was adjusted to yellow to aesthetic appeal. The x axis subtitle was also removed.

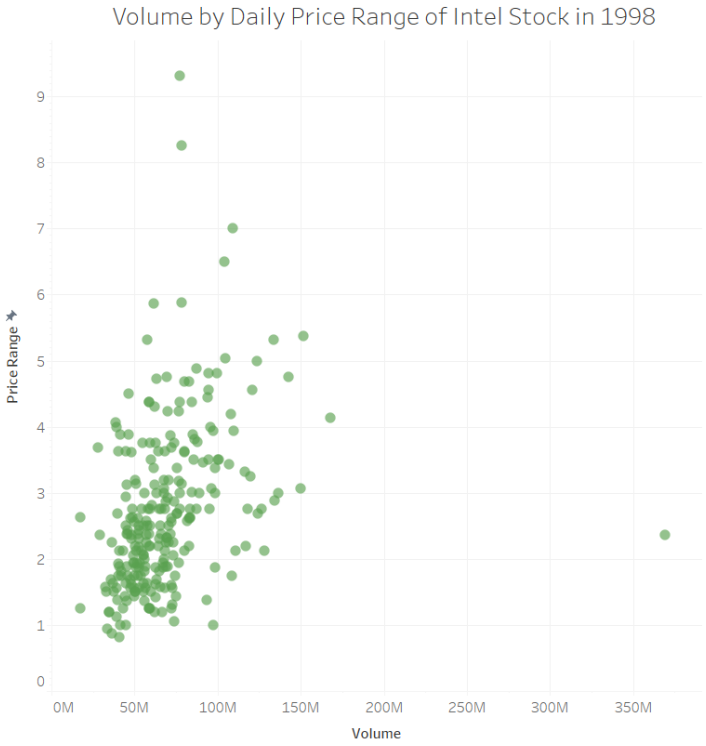


Problem 1: Trading Day

c. *Create a histogram of the daily stock Volume. In Tableau, the Histogram graph type in the Show Me box will be useful. Experiment with the bin size. In Tableau, after you have the histogram, right click the "Volume (bin)" in the data bar on the far left and use Edit. In Tableau, it's not the number of bins, but their width (in terms of data). You can set them that way in R as well with different parameters.*



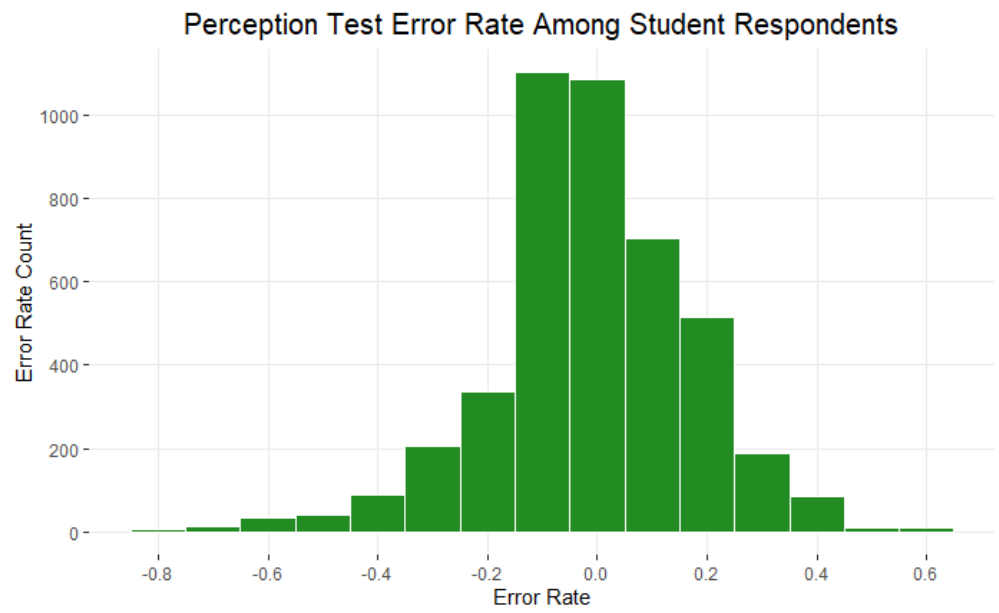
Steps: to create this graph the volume was moved to the column section a then see more was used to create a histogram. The bars were binned by 20 million intervals. The axis subtitle was removed, and the bars were changed to green.



d. *Create a scatterplot that graphs the Volume on the x-axis and the daily price range on the y-axis. You will need to create an additional column that contains the "range" of the prices for the day as the difference between the fields High and Low.*

Steps: This graph was created using the following steps. The data were accessed via the data source tab where a new field for the daily range was calculated using the original high variable. Once the data were prepped, the volume variable was moved to the rows field and the range variable to the column field. Both were set as average.

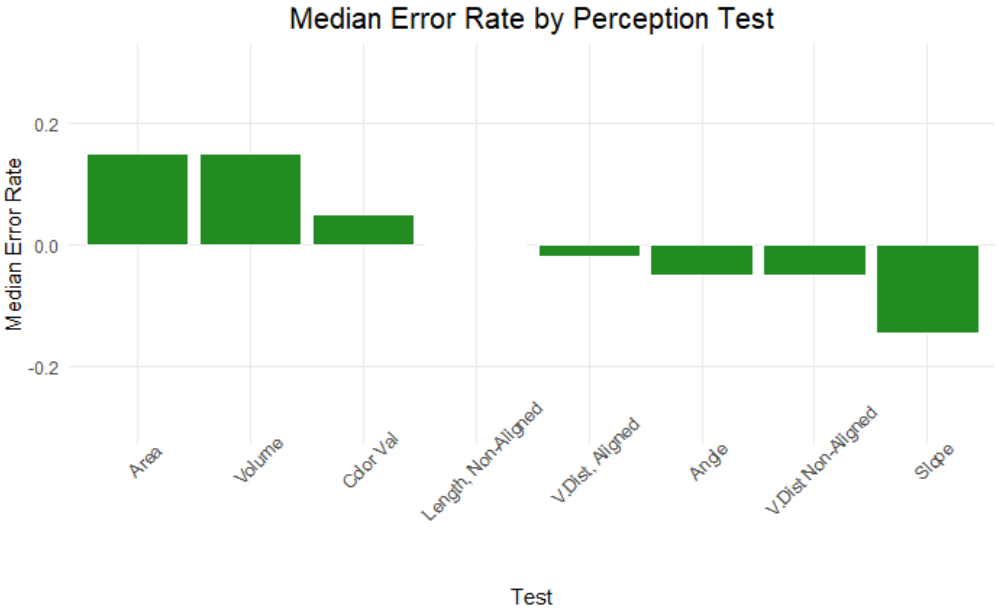
Problem 2: Perception Data



a. Create a histogram of the overall distribution of Error.

Interpretation: From this plot it can be inferred about one thousand responses were correct. It can also be observed students were more likely to underestimate the differences than overestimate the differences in the test.

Steps: To create this histogram I set the x y range and interval ticks so the graph would represent the range of error values without overwhelming the space. Then I created the graph using ggplot and geom_histogram. Once the graph was created, I began to edit the theme removing x var gridlines and softening the y gridlines.

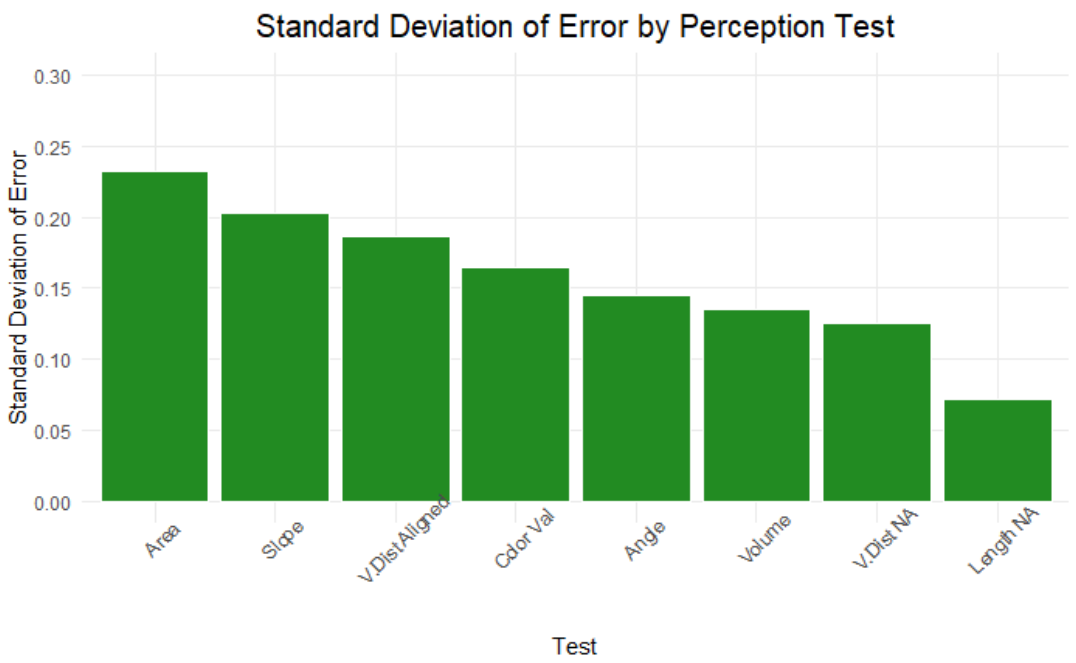


b. A bar graph of the median Error by Test (aka Error vs. Test). Do not subdivide by Display or the Trial. Order the x-axis to make the graph as clear as possible. Remember, for bar graphs in general, do not necessarily keep the default order (e.g. alphabetical) of the x-axis.

Interpretation: It can be inferred from the chart the perception was correctly interpreted by most students for the length non-aligned. Most students tend to overshoot the perceptual differences for the vertical distance aligned, angle , vertical distance nonaligned and slope. Alternatively, the area, volume and color value perception test had the highest median area rate meaning students tended to under estimate the perceptual differences in these tests.

Steps: For this bar graph the data was first subset to breakdown median error rate by test number. This subset was then used to create a bar graph using ggplot and geom_bar . The range and titles of the graphs were adjusted to display the data proportionately.

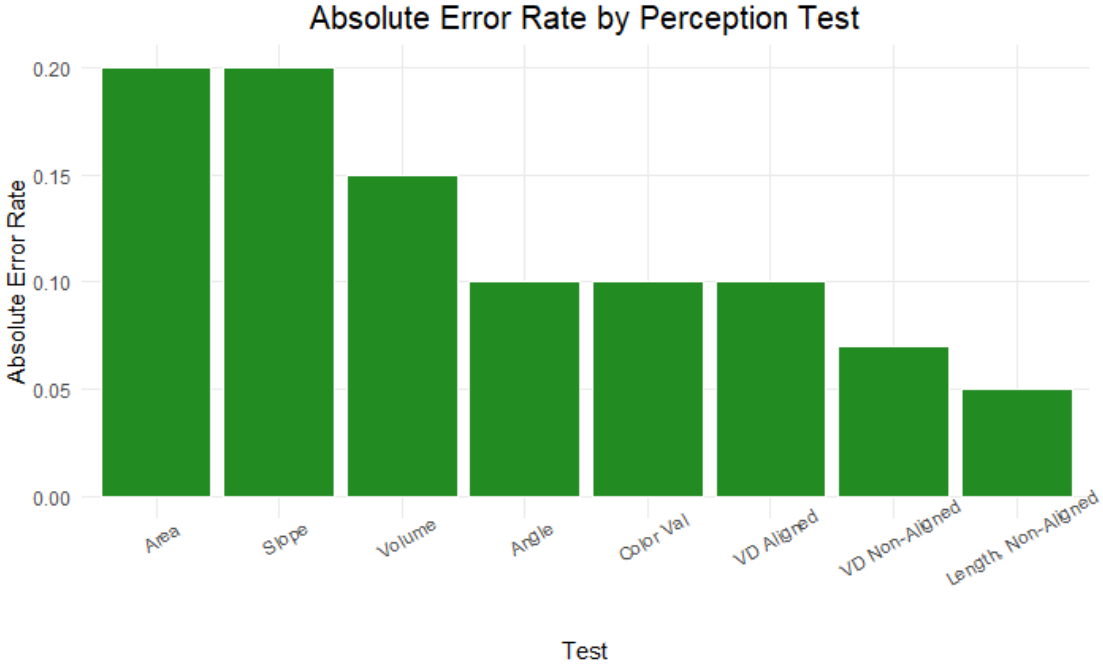
Problem 2: Perception Data



c. A bar graph of the standard deviation of the Error by Test. Remember that this measures the spread of how widely subjects varied in their responses. Again, order the x-axis to make the graph clear.

Steps: A subset for the standard deviation of each test was created. This data was plotted to a bar graph using ggplot. After the plot was created the theme and labels were adjusted. Finally the tick variables were created and then passed to the final graph.

Interpretation: Students had the most similar scores for the perception test, Length Non-Aligned, followed by Vertical Distance Not Aligned, Volume, Angle, color /Value, Vertical Distance Aligned, Slope, and finally Area. For those tests with a larger standard deviation it can be inferred student responses varied by a greater range.

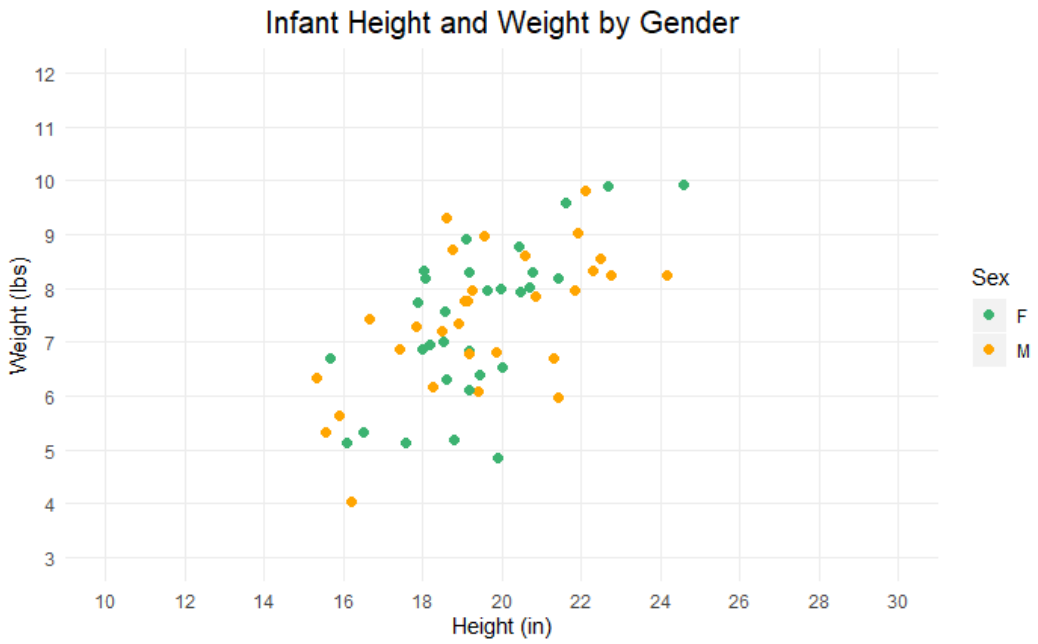


d. Create a new field called AbsoluteError by computing the absolute value of the Error field you created. Then do the same as in (b) with the AbsoluteError.

Steps: A subset of this data was created to represent the absolute error of the error for each test. Using this test a bar chart was created using ggplot. The theme and labels were adjusted, and new tick marks set.

Interpretation: From this chart we can see that most students responded incorrectly to the area and slope test, followed by the volume test and then the angle, color, and vertical distance aligned test which all had the same error rate. The Vertical Distance Aligned and the Length Non-Aligned Tests had the lowest error rate, meaning most students were better able to interpret the perceptual differences in those test.

Problem 3: Infant Data

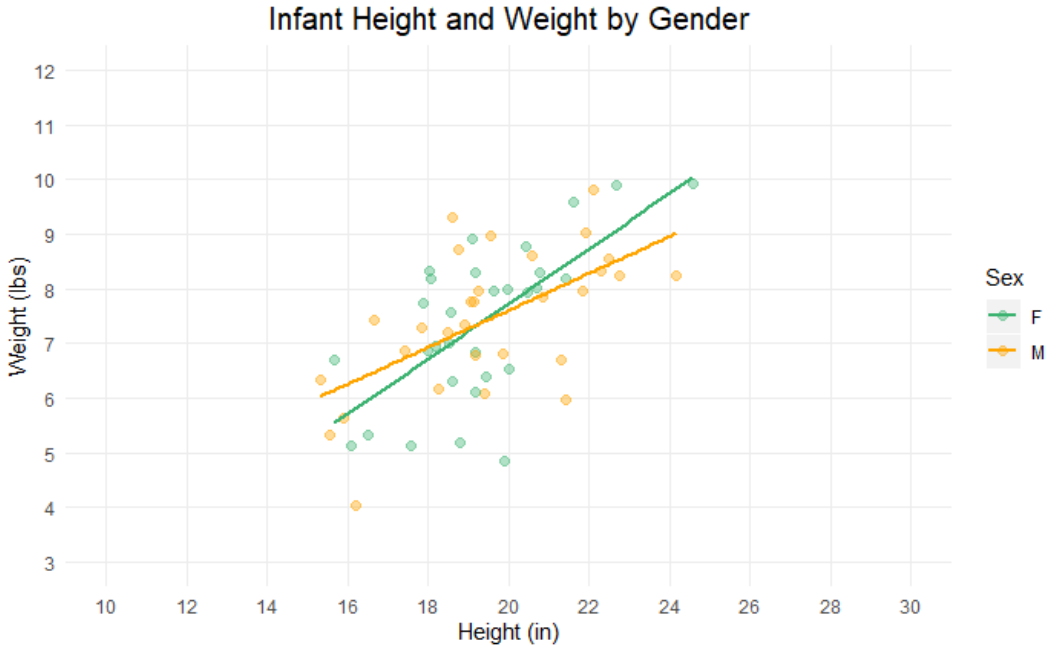


a. Graph the data as a scatter plot of Height.in on the x-axis and Weight.lbs on the y-axis. Differentiate in the plot between M or F values for Sex, but graph both on the same plot.

Steps: This scatter plot was created using ggplot. Specification were made to change the default data point colors and tick marks. After the data were plotted the plot themes and labels were changed.

b. Then create another single graph that has separate trend lines for the two populations on the graph. Adjust both the line and data-point weight and color to make the scatter plot and trend lines stand out.

Steps: This plot was created using ggplot with adjusted color specifications for the data points. The axis range was adjusted, and a range of tick marks was created to better display the data. A trend line was also added to this graph using stat smooth.

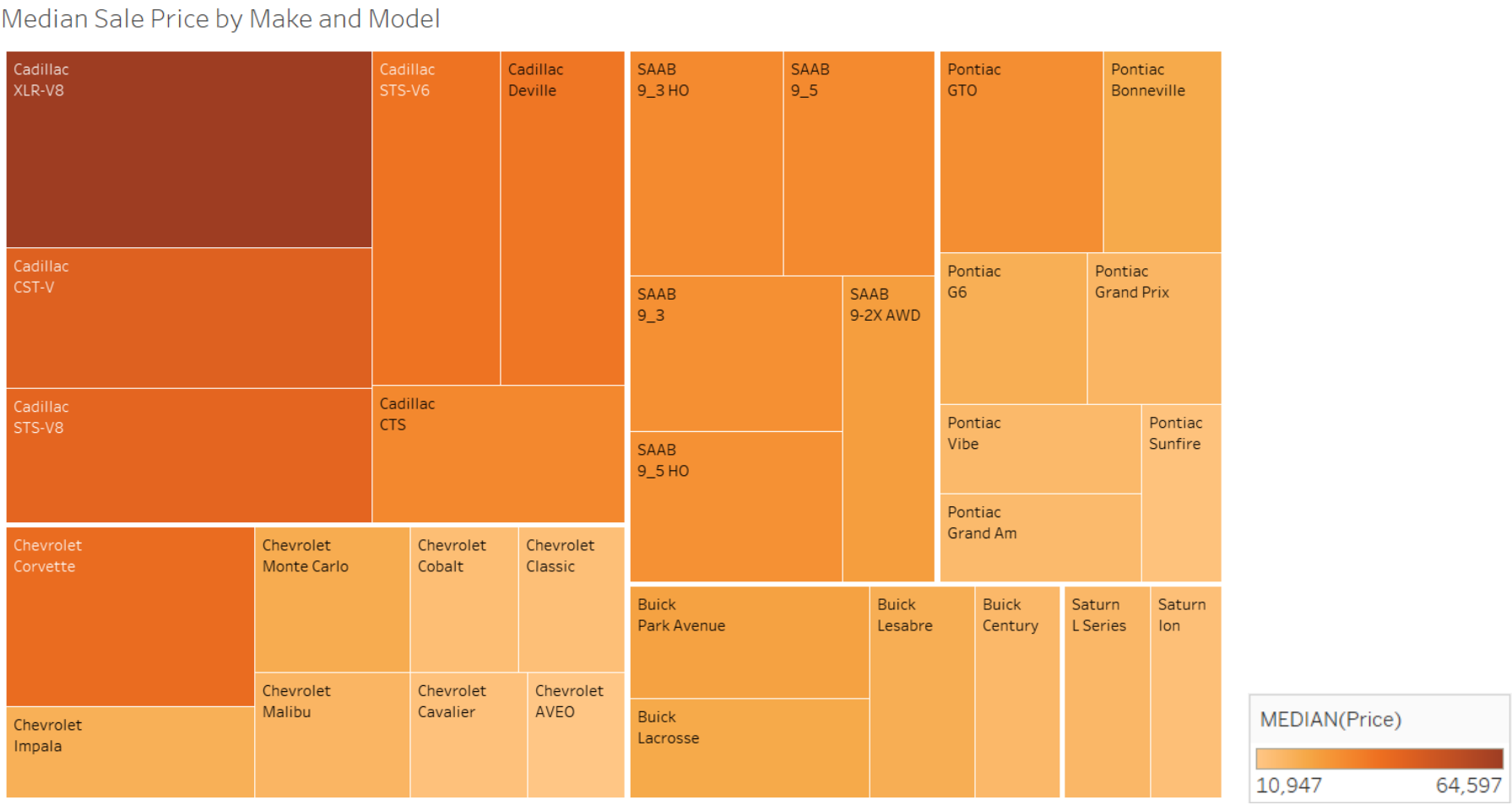


c. Explain in a short paragraph the decisions you made here and their impact on the visualization.

The infant data set has data points that clustered between 15-25 on the height and 4 and 10 in the weight variable. Since the data appeared clustered, I initially shortened the scale around those two ranges to observed if a new pattern might be observed. This change of scale did not show any new patterns so I chose a scale for both variables which would spread out the points a little but not dominate the entire plot. Color used to differentiate the sex is gender neutral to appeal to a modern audience. The background was set to white and the grid lines set to a very light grey to keep them as reference points that do not take away from the focus on the data.

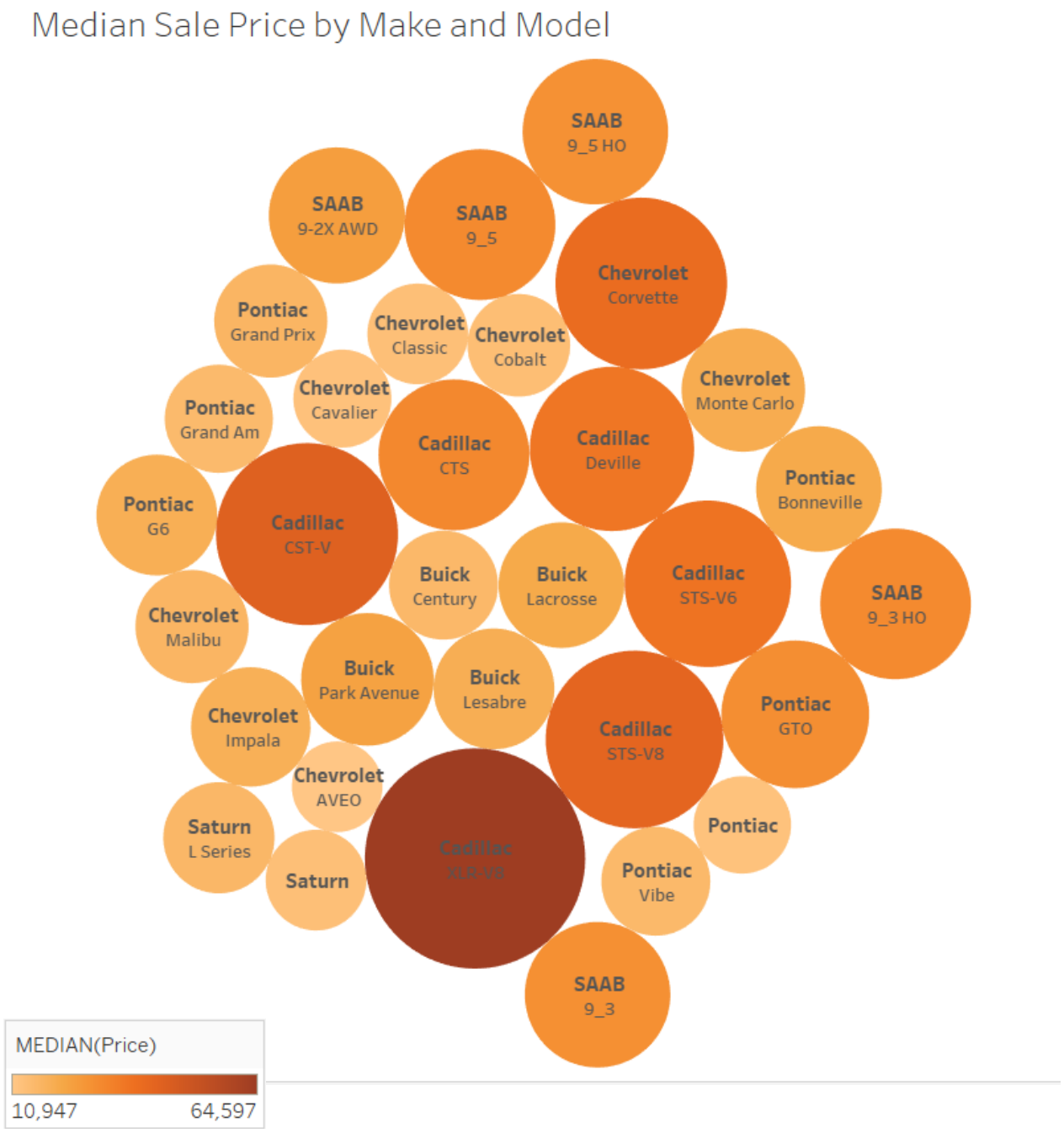
Problem 4: GM Car Price

a. A treemap based on Price with a main subdivision for the Make of the car and a minor subdivision based on the Model. Because each row of the data file represents a single car but each box in the treemap represents all the cars with a given make and model, pay very close attention to what kind of aggregation is being used.



Steps: To create this tree map I moved the make and model variables to the column section, selected tree map form the see more section then added average price as a detail layer for color and size and changed the color gradient to orange.

Problem 4: GM Car Price



b. Create a packed bubble chart of the same type

Steps: To create this bubble chart I moved the make and model variables to the column section, selected bubble chart from the see more section, added average price as a detail layer for color and size and changed the color gradient to orange.

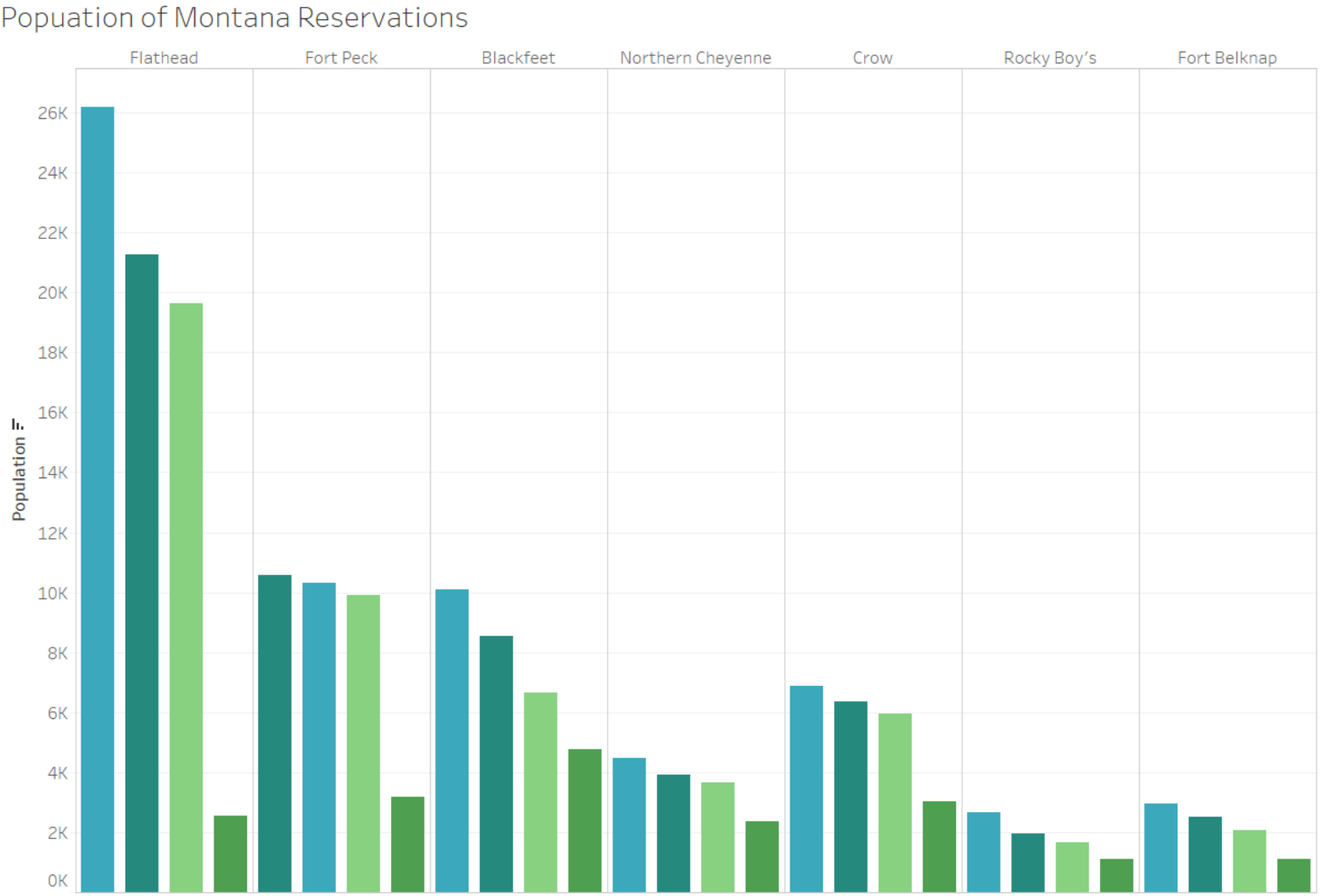
c. Write a short paragraph discussing the differences between the two plots. Describe for each something that displayed more clearly than with the other

In the treemap a single color gradient applied to all the cars and varied by median sale price clearly visualizes differences among makes and models and aids intra and inter make median sale price comparison. This is not easily done in the bubble chart. Applying a single gradient color to the bubble chart only varying by median sale price across make and models was not effective. Although the gradient could be set to vary by sale price across all makes and models uniformly this made intra make comparison very difficult considering models are not automatically clustered by make. Using the tree map it is very easy to identify both which car was the highest median sale price across all makes and what car has the highest median sale price by make.

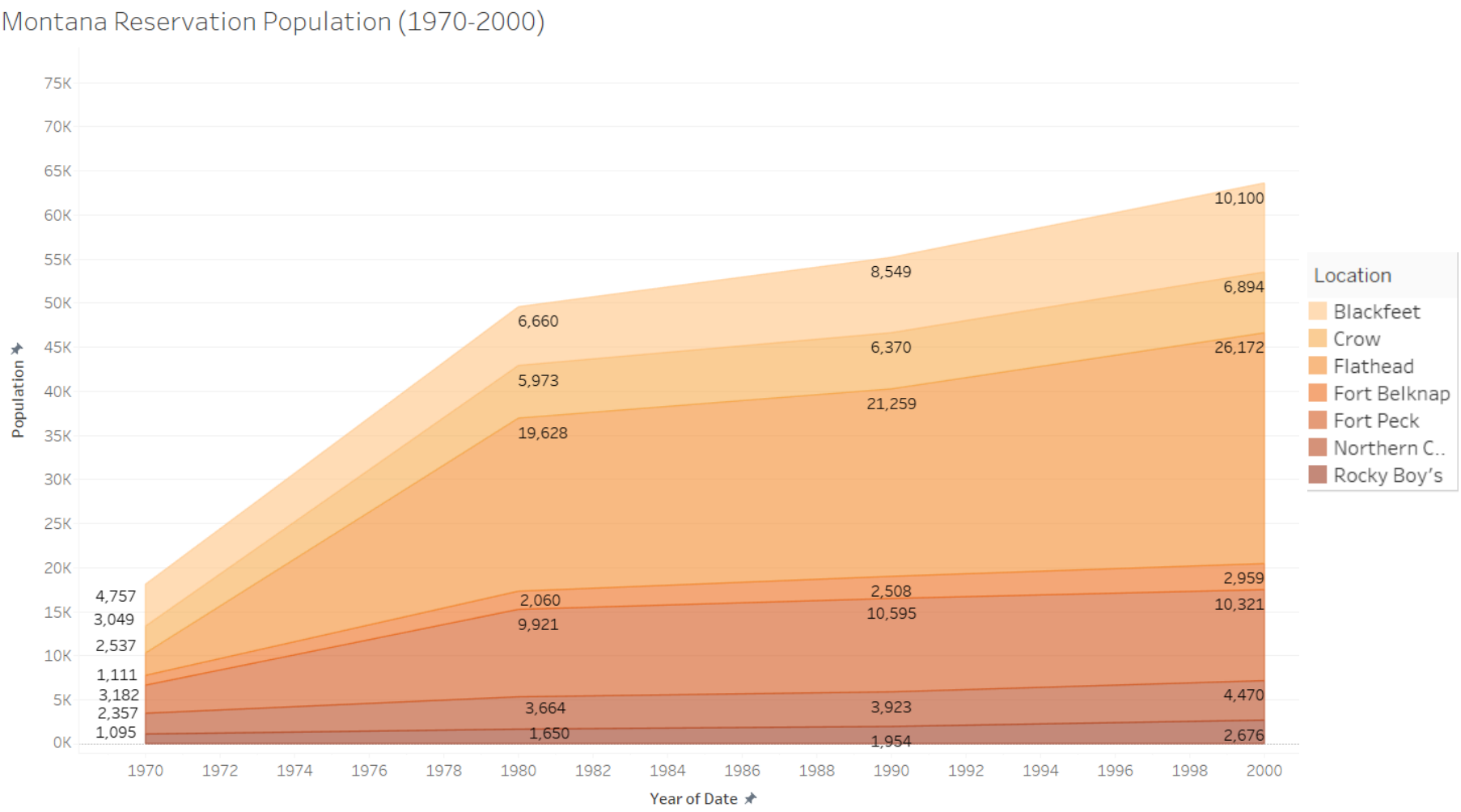
Although intra make comparison is difficult in the bubble chart, using the bubble chart which varies by median sale price single color gradient and size varying by the clearly displays the median sale price of most cars. Using the bubble graph we can clearly observe most cars sold have a price between 10 - 20k.

Problem 5: Reservation

a. One chart that graphs the population growth over the years for the individual reservations.



b. One that graphs the total reservation population subdivided among the different reservations for each year. The difference between this and (a) is that in (b) we are not looking only at each population individually but at the growth of the total population of all of them together, then subdivided by the reservations.



Steps: To create this chart location and sum population as sum were moved to the column and sum population was also moved to the row. Lines discrete was selected from the show me section and the color changed to orange.

c. One that graphs the population distribution over years for each reservation with a box-and-whisker plot. The 'distribution over years' means we are visualizing a distribution, i.e. multiple samples of something. In this case that is multiple samples of population value, one per year. For each reservation, we have four different year samples of population, so we will have a box-and-whiskers 'column' per reservation showing the distribution of population values at that location during this overall period.

Steps: To create this graph location and year as a date object were moved to the columns section. The sum population was also moved to rows. Then the box plot was selected from the see more function.

Question 2 R Code:

```
#setup

#load libs
library (ggplot2)
library(dplyr)
library(gcookbook)
library(plyr)

getwd()

#load and preview data
data <-read.csv("PerceptionExperiment.csv")
head(data,5)
#create error column
perception <-mutate(data,
                    error = Response - TrueValue)
summary(perception$error)
head(perception)

#DONE
#a. Create a histogram of the overall distribution of Error
ticks<-seq(0, 1400, by=200)
ticks2<-seq(-.8, .8, by=.2)

per_hist <-ggplot(perception, aes(x=error)) + geom_histogram(binwidth = .1, fill
='forestgreen', colour = 'white') + scale_y_continuous(breaks = ticks) +
scale_x_continuous(breaks = ticks2)
per_hist

#edit theme and add labels
per_hist +xlab('Error Rate') + ylab('Error Rate Count') + ggtitle('Perception Test Error
Rate Among Student Respondents') +
  theme(plot.title=element_text(size=15, hjust = 0.5),
        panel.grid.major.y = element_line(colour="gray92"),
        panel.grid.major.x = element_line(colour="gray92"),
        panel.background = element_rect(fill = "white")
        )
```

#b. Create a bar graph of the median Error by Test (aka Error vs. Test). Do not subdivide by Display or the Trial. Order the x-axis to make the graph as clear as possible. Remember, for bar graphs in general, do not necessarily keep the default order (e.g. alphabetical) of the x-axis.

```
#create median error var for each test (Ref R Graphing p.358)
data_error_test <-ddply(perception, 'Test', summarise, Median = median(error))
data_error_test
#rename tests
Test <-factor(c("Angle", "Color Value", "Slope","Vertical Distance, Non-Aligned","Area",
"Length, Non-Aligned", "Veritcal Distance, Aligned", "Volume"))
Test2 <- revalue(Test, c("Angle"="Angle", "Color Value" = "Color Val", "Slope" =
"Slope","Vertical Distance, Non-Aligned"="V.Dist Non-Aligned","Area" = "Area",
"Length, Non-Aligned" = "Length, Non-Aligned", "Veritcal Distance, Aligned" = "V.Dist,
Aligned", "Volume"="Volume") )

#create bar graph (Ref R Graphing p.20)
#tick_mederror<-seq(-.2, .25, by=.05)
#reorders cat vars on x , (x= reorder(Test,-Median)
error_test <- ggplot(data_error_test, aes(x= reorder(Test2,-Median), y=Median)) +
                geom_bar(stat="identity", fill ='forestgreen', colour = 'white' )  +
                scale_y_continuous( limits=c(-.3, .3))

#, breaks = tick_mederror)

#edit theme and add labels
MedErrorByTest <- error_test + xlab('Test') + ylab('Median Error Rate') +
ggtitle('Median Error Rate by Perception Test') +
  theme(plot.title=element_text(size=15, hjust = 0.5),
        panel.grid.major.y = element_line(colour="gray92"),
        panel.grid.major.x = element_line(colour="gray92"),
        panel.background = element_rect(fill = "white"),
        axis.ticks.y = element_blank(),
        axis.ticks.x = element_blank(),
        axis.text.x = element_text(angle = 45)
  )
MedErrorByTest
```

#c. Create a bar graph of the standard deviation of the Error by Test. Remember that this measures the spread of how widely subjects varied in their responses. Again, order the x-axis to make the graph clear.

```
#create sd of error variable for each test (Ref R Graphing p.358)
error_sd_bytest <-ddply(perception, 'Test', summarise, error_sd = sd(error))
error_sd_bytest

#rename tests
Test <-factor(c("Angle", "Color Value", "Slope","Vertical Distance, Non-Aligned","Area",
"Length, Non-Aligned", "Veritcal Distance, Aligned", "Volume"))
Test3 <- revalue(Test, c  ("Angle"="Angle", "Color Value" = "Color Val", "Slope" =
"Slope","Vertical Distance, Non-Aligned"="V.Dist NA","Area" = "Area", "Length, Non-
Aligned" = "Length NA", "Veritcal Distance, Aligned" = "V.Dist Aligned",
"Volume"="Volume") )

#create ticks
ticks_2c<-seq(0, .3, by=.05)

#create bar graph (Ref R Graphing p.20)
#tick_mederror<-seq(-.2, .25, by=.05)
#reorders cat vars on x , (x= reorder(Test,-Median)
error_sd_test <- ggplot(error_sd_bytest, aes(x= reorder(Test3,-error_sd), y=error_sd))+
                  geom_bar(stat="identity",fill ='forestgreen', colour = 'white')+
                  scale_y_continuous(breaks = ticks_2c,  limits=c(0, .3))
error_sd_test

#edit theme and add labels
error_sd_test_2C <- error_sd_test + xlab('Test') + ylab('Standard Deviation of Error') +
ggtitle('Standard Deviation of Error by Perception Test') +
  theme(plot.title=element_text(size=15, hjust = 0.5),
        panel.grid.major.y = element_line(colour="gray92"),
        panel.grid.major.x = element_line(colour="gray92"),
        panel.background = element_rect(fill = "white"),
        axis.ticks.y = element_blank(),
        axis.ticks.x = element_blank(),
        axis.text.x = element_text(angle = 45)
  )
error_sd_test_2C
```

#d. Create a new field called AbsoluteError by computing the absolute value of the Error field you created. Then do the same as in (b) with the AbsoluteError.

```
#create absolute error column
perception_2d <-mutate(perception,
                abs_error = abs(error))
summary(perception_2d$abs_error)
head(perception_2d)

#create abs error var for each test (Ref R Graphing p.358)
abserror_bytest_df <-ddply(perception_2d, 'Test', summarise, absolute_var =
median(abs_error))
abserror_bytest_df

#rename tests
Test <-factor(c("Angle", "Color Value", "Slope","Vertical Distance, Non-Aligned","Area",
"Length, Non-Aligned", "Veritcal Distance, Aligned", "Volume"))
Test2 <- revalue(Test, c("Angle"="Angle", "Color Value" = "Color Val", "Slope" =
"Slope","Vertical Distance, Non-Aligned"="VD Non-Aligned","Area" = "Area", "Length,
Non-Aligned" = "Length, Non-Aligned", "Veritcal Distance, Aligned" = "VD Aligned",
"Volume"="Volume") )

#create ticks
ticks_2d<-seq(0, .3, by=.05)

#create bar graph (Ref R Graphing p.20)
#tick_mederror<-seq(-.2, .25, by=.05)
#reorders cat vars on x , (x= reorder(Test,-Median)
test_abserror_bar <- ggplot(abserror_bytest_df, aes(x= reorder(Test2,-absolute_var),
y=absolute_var)) + geom_bar(stat="identity", fill ='forestgreen', colour = 'white' )
test_abserror_bar + scale_y_continuous(breaks = ticks_2d,limits=c(0, .3))
test_abserror_bar

#edit theme and add labels
test_abserror_bar2d <- test_abserror_bar + xlab('Test') + ylab('Absolute Error Rate') +
ggtitle('Absolute Error Rate by Perception Test') +
  theme(plot.title=element_text(size=15, hjust = 0.5),
        panel.grid.major.y = element_line(colour="gray92"),
        panel.grid.major.x = element_line(colour="gray92"),
        panel.background = element_rect(fill = "white"),
```

```
      axis.ticks.y = element_blank(),
      axis.ticks.x = element_blank(),
      axis.text.x = element_text(angle = 30)
  )
test_abserror_bar2d
```

Question 3 R Code:

```
library (ggplot2)
library(dplyr)
library(gcookbook)
library(plyr)

#check working directory
getwd()

#load and preview data
data <-read.csv("InfantData.csv")
head(data,5)

#DONE - maybe edt colors ect.
#a.Graph the data as a scatter plot of Height.in on the x-axis and Weight.lbs on the y-
axis.
#Differentiate in the plot between M or F values for Sex, but graph both on the same
plot.
#create tick marks
height_x<-seq(10,30, by=2)
weight_y<-seq(3, 12, by=1)

#REFERENCE: r cookbook p78
height_weight<- ggplot(data, aes(x=Height.in, y=Weight.lbs, colour = Sex))+
geom_point(size=2)+ scale_y_continuous(breaks= weight_y, limits=c(3,12)) +
scale_x_continuous(breaks= height_x,limits=c(10,30))+scale_colour_manual(values =
c("mediumseagreen","orange"))

#edit theme and add labels
infant_a <- height_weight + xlab('Height (in)') + ylab('Weight (lbs)') + ggtitle('Infant
Height and Weight by Gender') +
```

```
  theme(plot.title=element_text(size=15, hjust = 0.5),
        panel.grid.major.y = element_line(colour="gray93"),
        panel.grid.major.x = element_line(colour="gray93"),
        panel.background = element_rect(fill = "white"),
        axis.ticks.y = element_blank(),
        axis.ticks.x = element_blank()
        )
infant_a

#DONE
```

#b. Then create another single graph that has separate trend lines for the two populations on the #graph. Adjust both the line and data-point weight and color to make the scatter plot and #trend lines stand out.

```
#create tick marks
height_x<-seq(10,30, by=2)
weight_y<-seq(3, 12, by=1)

#REFERENCE: r cookbook p78
height_weight_b<- ggplot(data, aes(x=Height.in, y=Weight.lbs, colour = Sex))+
geom_point(size=2, alpha = .4)+ scale_y_continuous(breaks= weight_y, limits=c(3,12))
+ scale_x_continuous( breaks= height_x,limits=c(10,30)) + stat_smooth(method=lm,
se=FALSE) + scale_colour_manual(values = c("mediumseagreen","orange"))

#edit theme and add labels
infant_b <- height_weight_b + xlab('Height (in)') + ylab('Weight (lbs)') + ggtitle('Infant
Height and Weight by Gender') +
  theme(plot.title=element_text(size=15, hjust = 0.5),
        panel.grid.major.y = element_line(colour="gray93"),
        panel.grid.major.x = element_line(colour="gray93"),
        panel.background = element_rect(fill = "white"),
        axis.ticks.y = element_blank(),
        axis.ticks.x = element_blank()
        )
infant_b
```