VisualizationExploration

Kevin Letourneau

May 22, 2019

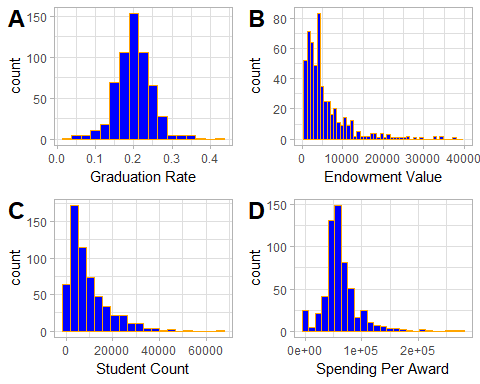
data <- read.csv(file="C:/Users/fuses/Downloads/institution\_data\_R.csv",header = TRUE)  
head(data)

## ï..institution\_name  
## 1 Alabama A&M University  
## 2 University of Alabama at Birmingham  
## 3 University of Alabama at Huntsville  
## 4 Alabama State University  
## 5 Auburn University at Montgomery  
## 6 Auburn University  
## basic grad\_rate  
## 1 Masters Colleges and Universities--larger programs 0.142  
## 2 Research Universities--very high research activity 0.209  
## 3 Research Universities--very high research activity 0.209  
## 4 Masters Colleges and Universities--larger programs 0.116  
## 5 Masters Colleges and Universities--larger programs 0.154  
## 6 Research Universities--high research activity 0.215  
## student\_count spending\_per\_award full\_time\_pct full\_time\_count  
## 1 4051 105331 0.938 3906  
## 2 11502 136546 0.727 10032  
## 3 5696 64418 0.744 5000  
## 4 5356 132407 0.910 5035  
## 5 4322 58541 0.694 3571  
## 6 19799 71999 0.910 19635  
## med\_sat\_value aid\_value endow\_value grad\_on\_time\_pct pell\_value  
## 1 823 7142 3808 0.10 0.712  
## 2 1146 6088 24136 0.29 0.351  
## 3 1180 6647 11502 0.16 0.328  
## 4 830 7256 13202 0.08 0.827  
## 5 970 4327 10736 0.09 0.401  
## 6 1215 8875 22092 0.37 0.169  
## fresh\_retain\_value full\_time\_fac\_pct EndowXSpend PellXSat RetainXSat  
## 1 0.631 0.828 401100448 585.976 519.313  
## 2 0.802 0.924 3295674256 402.246 919.092  
## 3 0.810 0.655 740935836 387.040 955.800  
## 4 0.622 0.670 1748037214 686.410 516.260  
## 5 0.632 0.601 628496176 388.970 613.040  
## 6 0.895 0.887 1590601908 205.335 1087.425  
## AidXSat AidXEndow id  
## 1 5877866 27196736 1  
## 2 6976848 146939968 2  
## 3 7843460 76453794 3  
## 4 6022480 95793712 4  
## 5 4197190 46454672 5  
## 6 10783125 196066500 6

grad <- ggplot(data = data, aes(x=grad\_rate)) +  
 geom\_histogram(binwidth = 0.025, bins = 20, fill = "blue", color = "orange")+  
 labs(x="Graduation Rate")+  
 theme\_light()  
  
  
  
endow <- ggplot(data = data, aes(x=endow\_value)) +  
 geom\_histogram(bins = 50, fill = "blue", color = "orange")+  
 labs(x="Endowment Value")+  
 xlim(0,40000)+  
 theme\_light()  
  
  
  
student <- ggplot(data = data, aes(x=student\_count)) +  
 geom\_histogram(bins = 20, fill = "blue", color = "orange")+  
 labs(x="Student Count")+  
 theme\_light()  
  
  
  
spending <- ggplot(data = data, aes(x=spending\_per\_award)) +  
 geom\_histogram(bins = 25, fill = "blue", color = "orange")+  
 labs(x="Spending Per Award")+  
 theme\_light()  
  
  
  
cowplot::plot\_grid(grad,endow,student,spending,  
 labels = "AUTO",  
 label\_size = 18,  
 align = "hv")

## Warning: Removed 11 rows containing non-finite values (stat\_bin).

## Warning: Removed 2 rows containing missing values (geom\_bar).



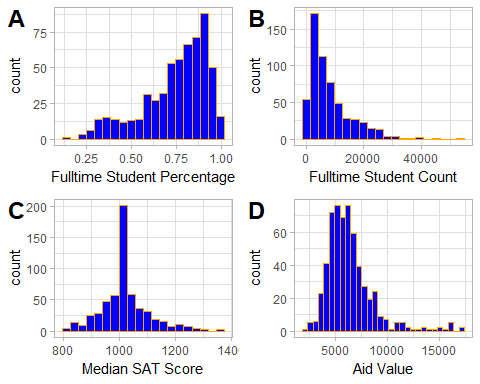
A: Grad\_rate has a strong bell shape curve with high center near its mean with a very even spread. This seems to have a normal distribution. From our SD we can see that we have 4 outliers on our lower threshold below 0.056, and we have 5 outliers on the upper threshold above 0.343.

B: Endow\_value has a large positive skew above our threshold shifting the center and mean well to the left side of this distribution making for a very long right tail. The spread is very uneven due to the skew. This is not a normal distribution. We can see that we likely have some extreme outliers on the high end of the distribution. This variable also has a significantly high variance meaning we have a large spread for our data. From our SD we can see that we have 0 outliers on our lower threshold below -33,686.836, and we have 7 outliers on the upper threshold above 47,391.595.

C: Student\_count has a large positive skew above our threshold shifting the center and mean well to the left side of this distribution making for a very long right tail. The spread is very uneven due to the skew. This is not a normal distribution. This variable also has a significantly high variance meaning we have a large spread for our data. From our SD we can see that we have 0 outliers on our lower threshold below -16520.903 as we can’t have less than 0 students, and we have 8 outliers on the upper threshold above 35,299.394.

D: Spending\_per\_award seems to have an odd distribution. It has a large positive skew above our threshold shifting the center and mean well to the left side of this distribution making for a very long right tail. The spread is very uneven due to the skew. This is not a normal distribution. This variable also has a significantly high variance meaning we have a large spread for our data. From our SD we can see that we have 0 outliers on our lower threshold below -38,911.013, and we have 9 outliers on the upper threshold above 162,893.814.

fullpct <- ggplot(data = data, aes(x=full\_time\_pct)) +  
 geom\_histogram(bins = 20, fill = "blue", color = "orange")+  
 labs(x="Fulltime Student Percentage")+  
 theme\_light()  
  
  
fullcount <- ggplot(data = data, aes(x=full\_time\_count)) +  
 geom\_histogram(bins = 20, fill = "blue", color = "orange")+  
 labs(x="Fulltime Student Count")+  
 theme\_light()  
  
  
medsat <- ggplot(data = data, aes(x=med\_sat\_value)) +  
 geom\_histogram(bins = 20, fill = "blue", color = "orange")+  
 labs(x="Median SAT Score")+  
 theme\_light()  
  
  
  
aid <- ggplot(data = data, aes(x=aid\_value)) +  
 geom\_histogram(bins = 30, fill = "blue", color = "orange")+  
 labs(x="Aid Value")+  
 theme\_light()  
  
  
cowplot::plot\_grid(fullpct, fullcount, medsat, aid,  
 labels = "AUTO",  
 label\_size = 18,  
 align = "hv")



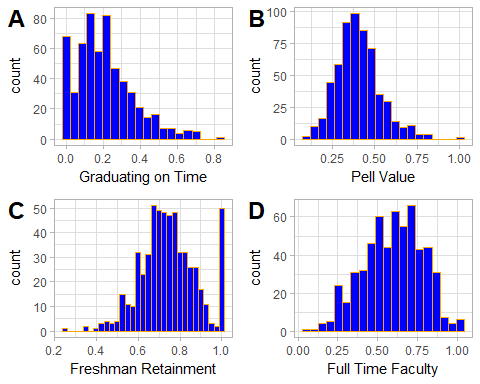
A: Full\_time\_pct has a negative skew just below our threshold shifting the center and mean to the right side of this distribution making for a long left tail. While it is below our threshold the spread is very uneven due to the skew. This is not a normal distribution. From our SD we can see that we have 1 outlier on our lower threshold below 0.201, and we have 0 outliers on the upper threshold above 1.291 as it is above 100%.

B: Full\_time\_count has a large positive skew above our threshold shifting the center and mean well to the left side of this distribution making for a very long right tail. The spread is very uneven due to the skew. This is not a normal distribution. This variable also has a significantly high variance meaning we have a large spread for our data. From our SD we can see that we have 0 outliers on our lower threshold below -14198.547, and we have 9 outliers on the upper threshold above 30453.925.

C: Med\_sat\_value has another odd distribution however this one is odd in a different way. We have an extremely high center and frequency at our mean with a fairly even spread from the mean. This distribution it a normal distribution however the extremely high center could be a cause for concern as it may mean that this variable is of little influence as most of its data is at one value. From our SD we can see that we have 0 outliers on our lower threshold below 762.38, and we have 6 outliers on the upper threshold above 1277.988.

D: Aid\_Value seems to have an odd distribution. It has a large positive skew above our threshold shifting the center and mean well to the left side of this distribution making for a very long right tail. The spread is very uneven due to the skew. This is not a normal distribution. This variable also has a significantly high variance meaning we have a large spread for our data. From our SD we can see that we have 0 outliers on our lower threshold below -368.154, and we have 14 outliers on the upper threshold above 13055.759.

gradtime <- ggplot(data = data, aes(x=grad\_on\_time\_pct)) +  
 geom\_histogram(bins = 20, fill = "blue", color = "orange")+  
 labs(x="Graduating on Time")+  
 theme\_light()  
  
  
  
pell <- ggplot(data = data, aes(x=pell\_value)) +  
 geom\_histogram(bins = 20, fill = "blue", color = "orange")+  
 labs(x="Pell Value")+  
 theme\_light()  
  
  
  
fresh <- ggplot(data = data, aes(x=fresh\_retain\_value)) +  
 geom\_histogram(binwidth = 0.025, fill = "blue", color = "orange")+  
 labs(x="Freshman Retainment")+  
 theme\_light()  
  
  
  
fullfac <- ggplot(data = data, aes(x=full\_time\_fac\_pct)) +  
 geom\_histogram(bins = 20, fill = "blue", color = "orange")+  
 labs(x="Full Time Faculty")+  
 theme\_light()  
  
  
cowplot::plot\_grid(gradtime, pell, fresh, fullfac,  
 labels = "AUTO",  
 label\_size = 18,  
 align = "hv")



A: Grad\_on\_time\_pct has a large positive skew above our threshold shifting the center and mean well to the left side of this distribution making for a very long right tail. The spread is very uneven due to the skew. This is not a normal distribution. From our SD we can see that we have 0 outliers on our lower threshold below -.26, and we have 6 outliers on the upper threshold above 0.675.

B: Pell\_Value is a near normal distribution with a decent bell shaped curve with some outliers on the higher end of the range around 0.8-1.0 that causes the range to be increased. There is a slight positive skewness however, it is not above or below our skew threshold of (-1.0, 1.0) therefor we will maintain that it is a near normal distribution with a high center near its mean and fairly even spread from its mean. From our SD we can see that we have 0 outliers on our lower threshold below 0.014, and we have 4 outliers on the upper threshold above 0.794.

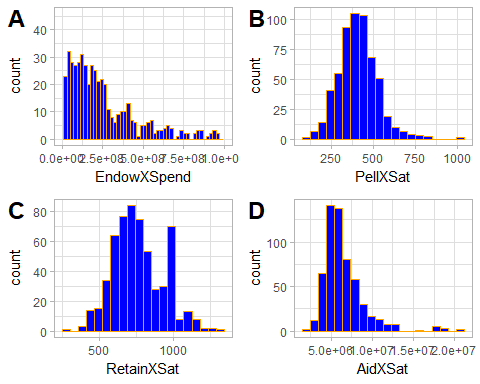
C: Fresh\_retain\_value has a decent bell shape curve with high center near its mean with a very even spread however, there are some possible outliers on the lower end of the spectrum causing the range of the histogram to be increased. Overall this does not seem to effect the distribution as nearly all of its data falls within the bell shaped area of the distribution. From our SD we can see that we have 1 outlier on our lower threshold below 0.351, and we have 0 outliers on the upper threshold as our threshold is above 100%.

D: Full\_time\_fac\_pct has a slight left skew showing that a higher portion of the data occurs around the 60%-80% mark right above the mean of 59.9%. The center is shifted slightly to the right with a fairly even spread. This is a near normal distribution. From our SD we can see that we have 1 outlier on our lower threshold below 0.0524, and we have 0 outliers on the upper threshold as our threshold is above 100%.

endowspend <- ggplot(data = data, aes(x=EndowXSpend)) +  
 geom\_histogram(bins = 50, fill = "blue", color = "orange")+  
 xlim(0,1000000000)+  
 theme\_light()  
  
  
  
pellsat <- ggplot(data = data, aes(x=PellXSat)) +  
 geom\_histogram(bins = 20, fill = "blue", color = "orange")+  
 theme\_light()  
  
  
  
retainsat <- ggplot(data = data, aes(x=RetainXSat)) +  
 geom\_histogram(bins = 20, fill = "blue", color = "orange")+  
 theme\_light()  
  
  
  
aidsat <- ggplot(data = data, aes(x=AidXSat)) +  
 geom\_histogram(bins = 20, fill = "blue", color = "orange")+  
 theme\_light()  
  
  
cowplot::plot\_grid(endowspend, pellsat, retainsat, aidsat,  
 labels = "AUTO",  
 label\_size = 18,  
 align = "hv")

## Warning: Removed 56 rows containing non-finite values (stat\_bin).

## Warning: Removed 2 rows containing missing values (geom\_bar).



A: EndowXSpending has a large positive skew above our threshold shifting the center and mean well to the left side of this distribution making for a very long right tail. The spread is very uneven due to the skew. This is not a normal distribution. This variable also has a significantly high variance meaning we have a large spread for our data. From our SD we can see that we have 0 outliers on our lower threshold below -3,850,369,898, and we have 10 outliers on the upper threshold above 4,869,272,108.

B: PellXSat has a strong bell shape curve with high center near its mean with a very even spread however, there are some possible outliers on the high end of the spectrum causing the range of the histogram to be increased. There is also a slight skew, but not above our threshold. Overall this does not seem to effect the distribution. From our SD we can see that we have 0 outliers on our lower threshold below 61.134, and we have 6 outliers on the upper threshold above 750.67.

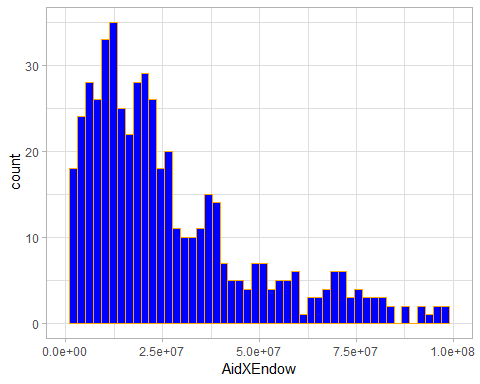
C: RetainXSat has a strong bell shape curve with high center near its mean with a very even spread. This seems to have a normal distribution. This variable has a neutral skewness we will assume normal distribution. From our SD we can see that we have 0 outliers on our lower threshold below 246.45, and we have 2 outliers on the upper threshold above 1285.57.

D: AidXSat has a large positive skew above our threshold shifting the center and mean well to the left side of this distribution making for a very long right tail. The spread is very uneven due to the skew. This is not a normal distribution. This variable also has a significantly high variance meaning we have a large spread for our data. From our SD we can see that we have 0 outliers on our lower threshold below -1,160,555.356, and we have 11 outliers on the upper threshold above 14,172,749.45.

aidendow <- ggplot(data = data, aes(x=AidXEndow)) +  
 geom\_histogram(bins = 50, fill = "blue", color = "orange")+  
 xlim(0,100000000)+  
 theme\_light()  
  
aidendow

## Warning: Removed 52 rows containing non-finite values (stat\_bin).

## Warning: Removed 2 rows containing missing values (geom\_bar).



A: AidXEndow has a large positive skew above our threshold shifting the center and mean well to the left side of this distribution making for a very long right tail. The spread is very uneven due to the skew. This is not a normal distribution. This variable also has a significantly high variance meaning we have a large spread for our data. We can see from out SD that we have 0 lower outliers on our lower threshold below -357,016,265.2, we do have 8 outliers at our upper threshold of 460,124,632.4.