Data analysis

Assignment on Descriptive Statistics

1. **Univariate data.**

**(a)  Construct a stem-and-leaf and histogram. Impose the empirical density estimate on the histogram. Discuss the results focusing on the shape of the plots and number of modes.**

*Code:*

*> stem(Credit$Limit)*

*> hist(Credit$Limit)*

*> ggplot(Credit, aes(Limit)) + geom\_histogram(aes(y = ..density..), bins=20, color="black", fill="white") + geom\_density(color="red", size=1.2) + theme\_bw()*

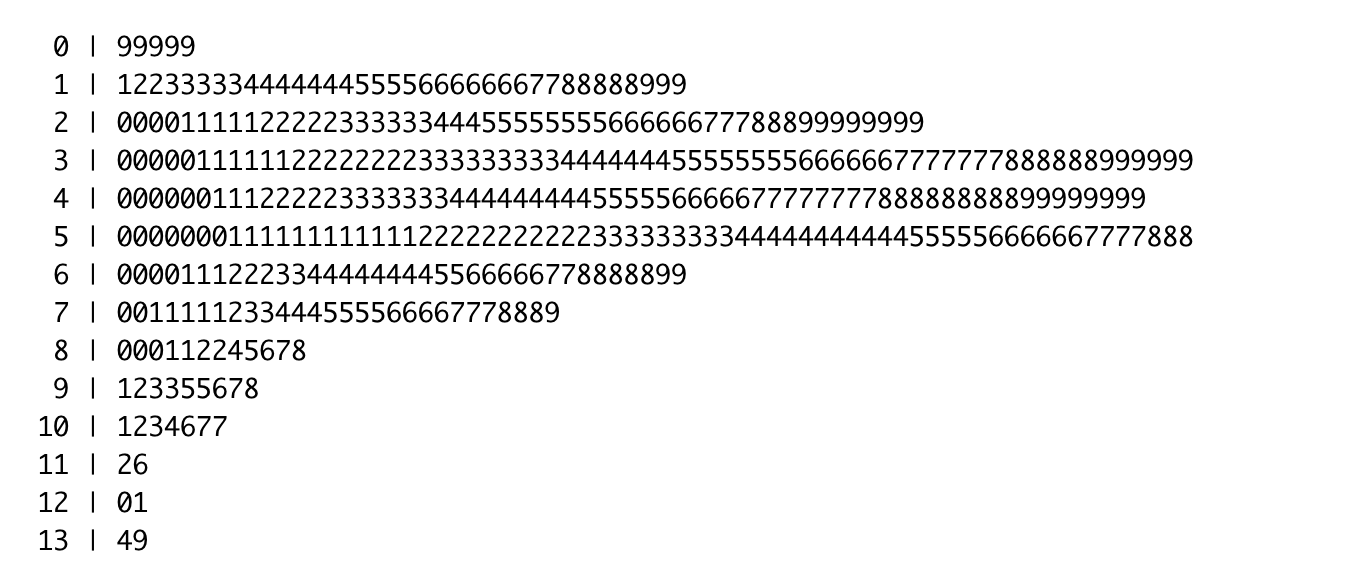


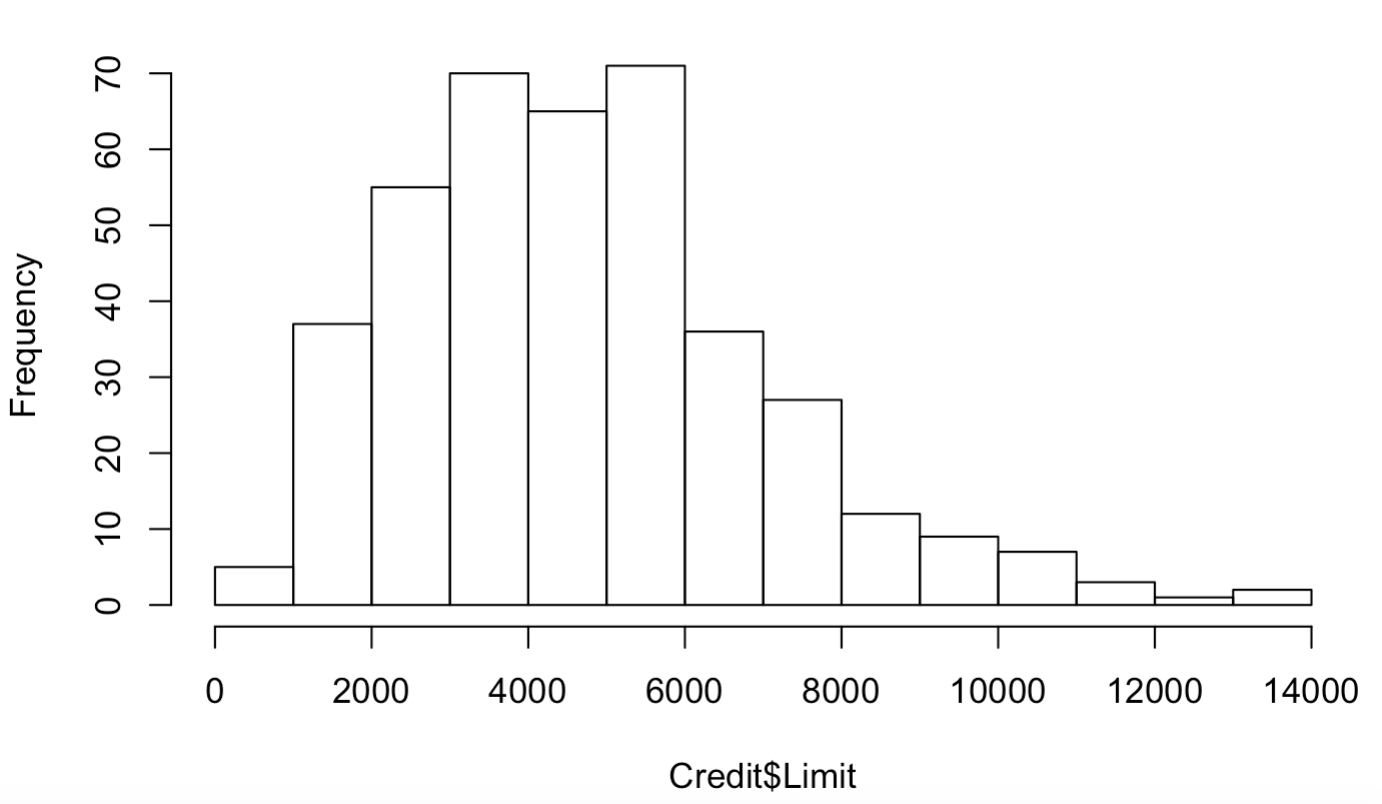
Figure 1:

Stem-and-leaf Plot

Steam-and-leaf plot let us quickly estimate min and max value of the data, shape of distribution. This chart doesn’t allow to get the exact values, because with three or more characters numbers are rounded.

Figure 2:

Histogram



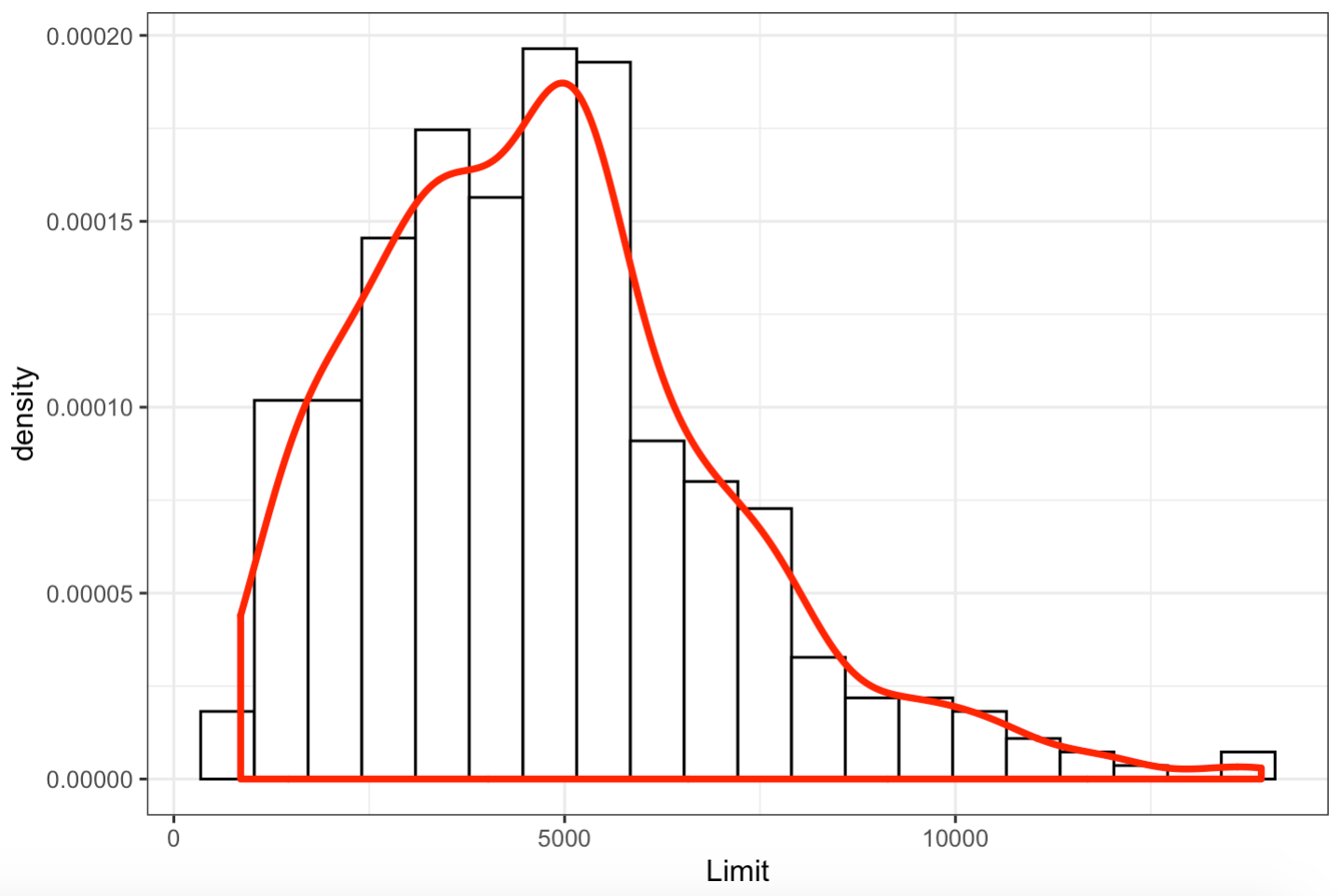


Figure 3:

Histogram with the imposed empirical density estimate

The density estimate clearly shows one peak in this data set. It is layered on top of the histogram plotted with total area 1.

**Conclusion:** So here we have a unimodal right-skewed distribution. There is no symmetry because of a long right tail.

**(b) Compute the mean and median. Based solely on that, conclude whether the distribution is skewed. Find the proportion of the data which are less than the mean value.**

*Code:*

*> x = Credit$Limit*

*> mean(x)*

***[1] 4735.6***

*> median(x)*

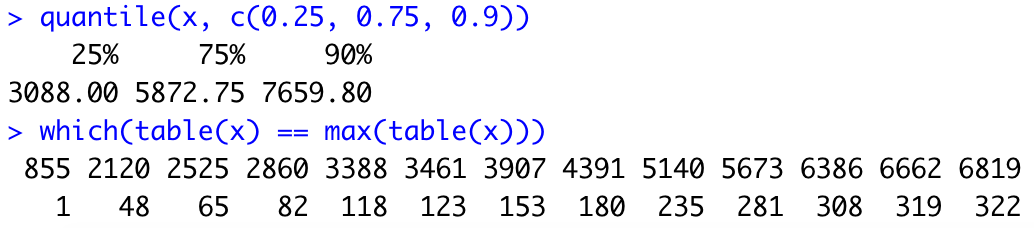
***[1] 4622.5***

*> sum(x <= mean(x))/length(x)\*100*

***[1] 52***

**Conclusion:** As we can see, median value is smaller than the mean. That means that distribution is right-skewed. The last line shows that 52% of the values are less than or equal to the sample mean.

**(c)  Compute the 1st and 3rd quartiles, the 90th quantile and the mode. Explain the meaning of the obtained quantities. Find the value that cuts off the top 25% of the data.**

*Code:*

**Conclusion:**

* 1st quartile is the value for which 25% of the observations are smaller and 75% are larger. Here it is **3088**
* 3rd quartile is the value for which 75% of the observations are smaller and 25% are larger. It means that the value **5872.75** cuts off the top 25% of the data.
* 90% quantile s the value for which 90% of the observations are smaller and 10% are larger. Here it is **7659.80**
* Mode is a most frequent value in the dataset. In my example there are **13** most frequent values. Each of them occurs **twice**

**(d)  Compute the range, the sample standard deviation and the IQR. Construct the boxplot of the data. Comment on the boxplot including skewness, outliers etc.**

*Code:*

*> diff(range(x))*

***[1] 13058***

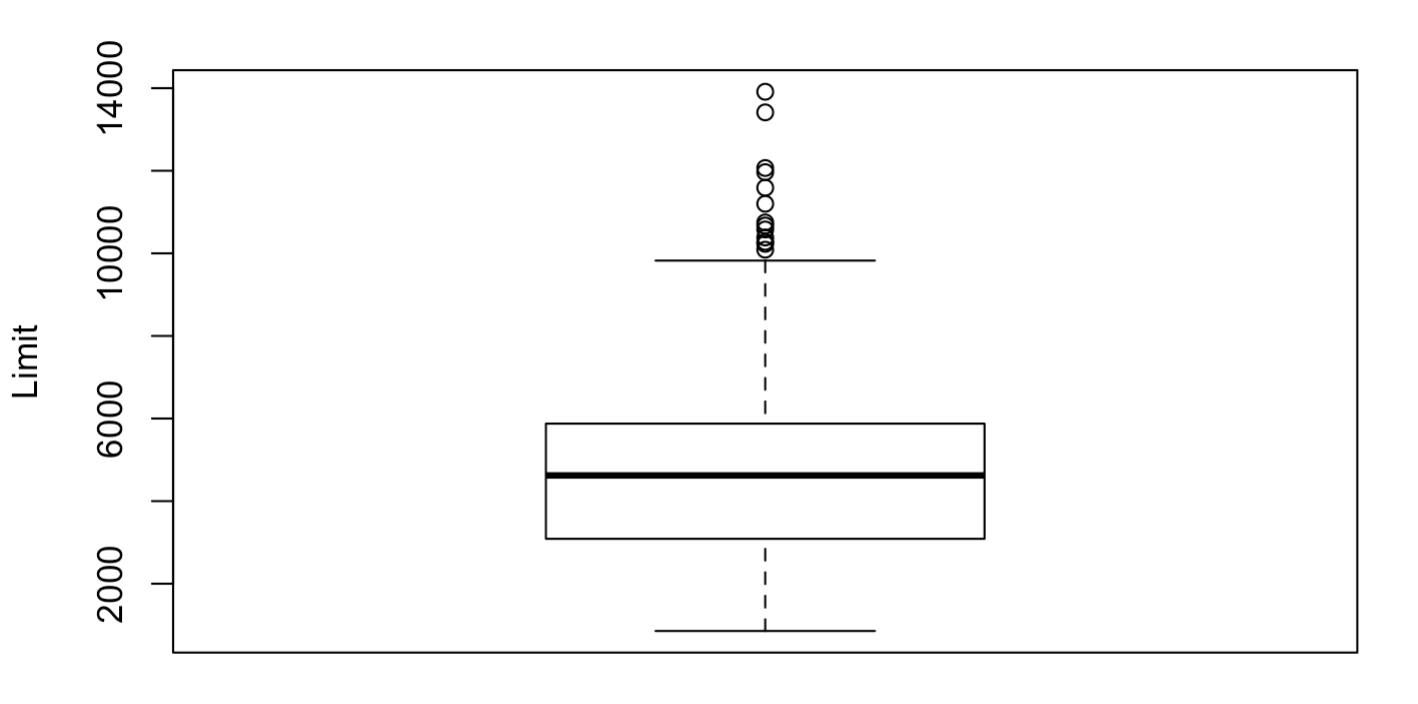
*> sd(x)*

***[1] 2308.199***

*> IQR(x)*

***[1] 2784.75***

*> boxplot(x, ylab="Limit")*

Figure 4: Boxplot

**Conclusion:**

* Boxplot shows that the dataset is skewed right. It has a long right tail and short left tail.
* Median separates the interquartile range into two quite symmetric parts. So data between 1st and 3rd quartile is distributed closed to normal
* The length of the upper whisker is bigger then the lower. Minimum value is closer to the center
* There are several outliers after the upper hinge, so the sample range is not very useful for measuring variation here.

**(e) Check whether the empirical distribution is normal by examining the QQ-plot.**

***Code:***

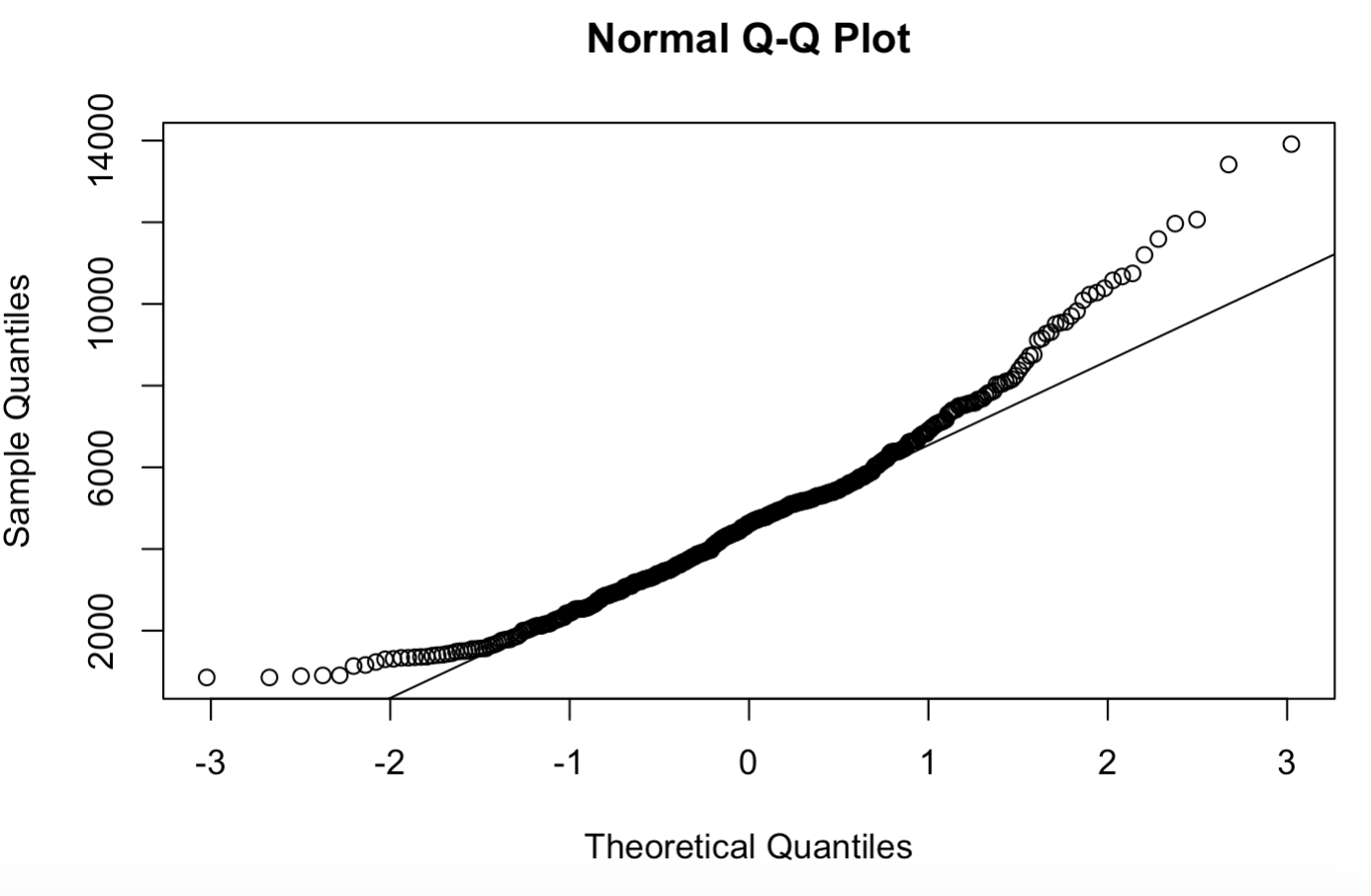
*> qqnorm(x)*

*> qqline(x)*

Figure 5: QQ-plot

**Conclusion:**

* The graph curves up, as this data has a long right tail.
* Distribution is closed to normal only at the central part between 1st and 3rd quartile as it looks like a straight line

******

1. **Bivariate data.**

**(a)  Create side-by-side boxplots. Compare the centers and spreads.**

***Code:***

*> boxplot(States$SATV, States$SATM, names = c("SATV", «SATM»))*

*> plot(density(States$SATV), main="densityplots of SATV and SATM")*

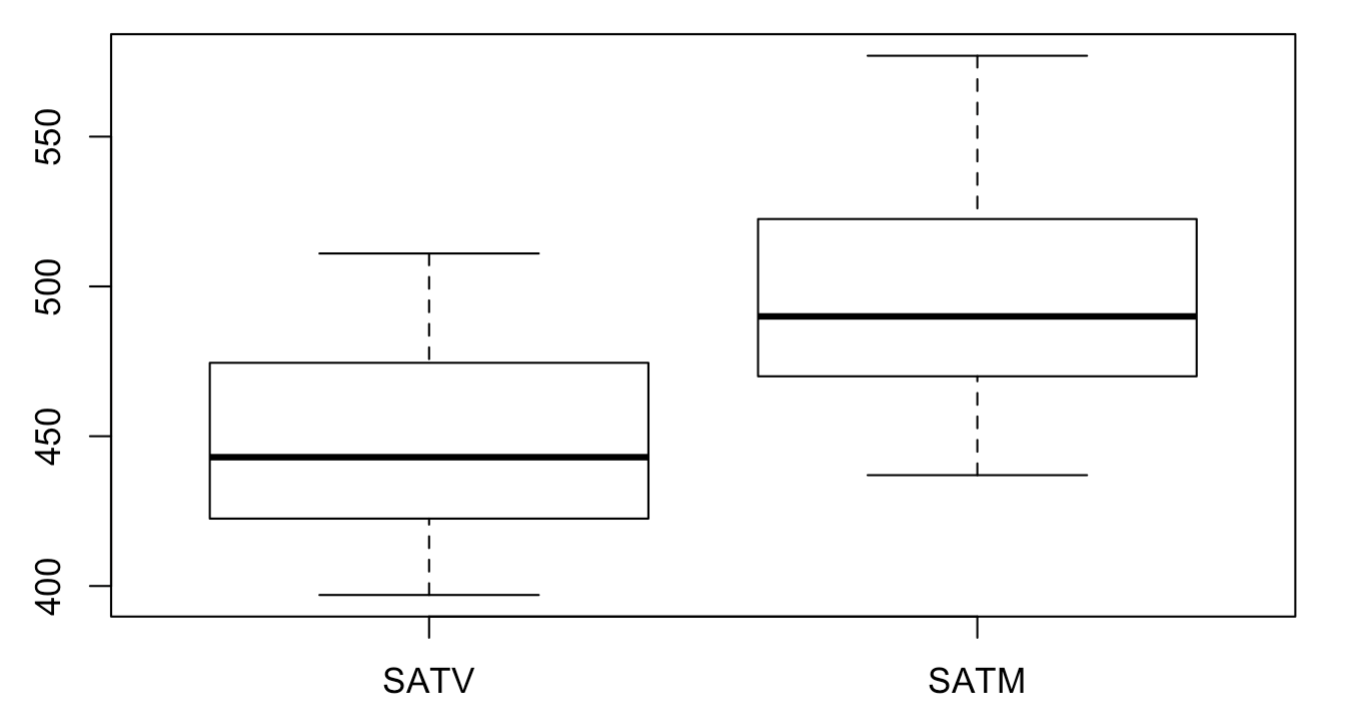
*> lines(density(States$SATM), lty=2)*

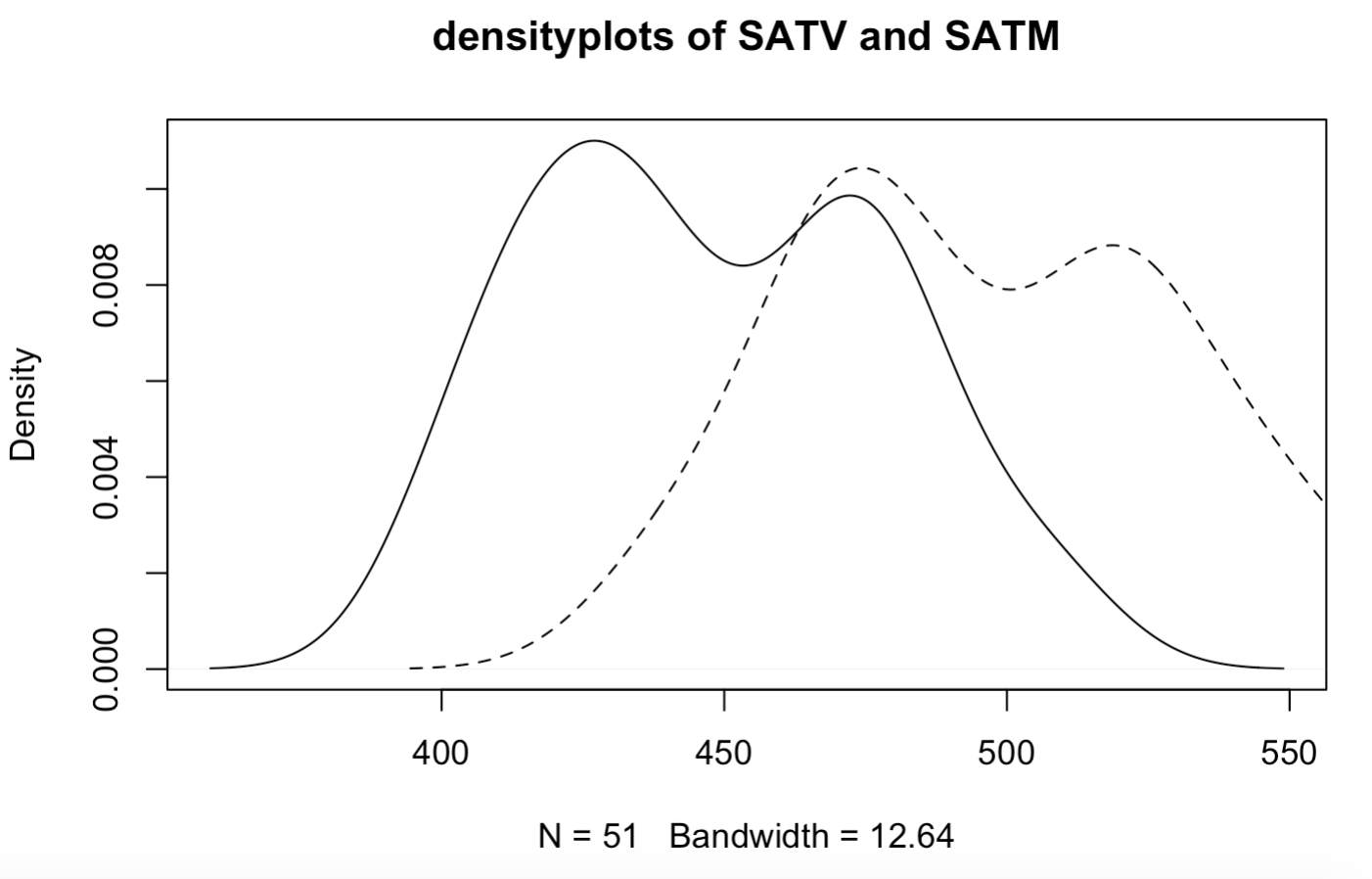
Figure 6. Comparing boxplots

Figure 7. Densityplots

**Conclusion:**

* Median of «*math* component of the Scholastic Aptitude Test» is bigger than «*verbal* component of the Scholastic Aptitude Test».
* Distributions looks quite similar, with different centers
* Both datasets have no outliers

**(b)  Draw the scatter plot. Comment on the possible dependence and presence of outliers.**

***Code:***

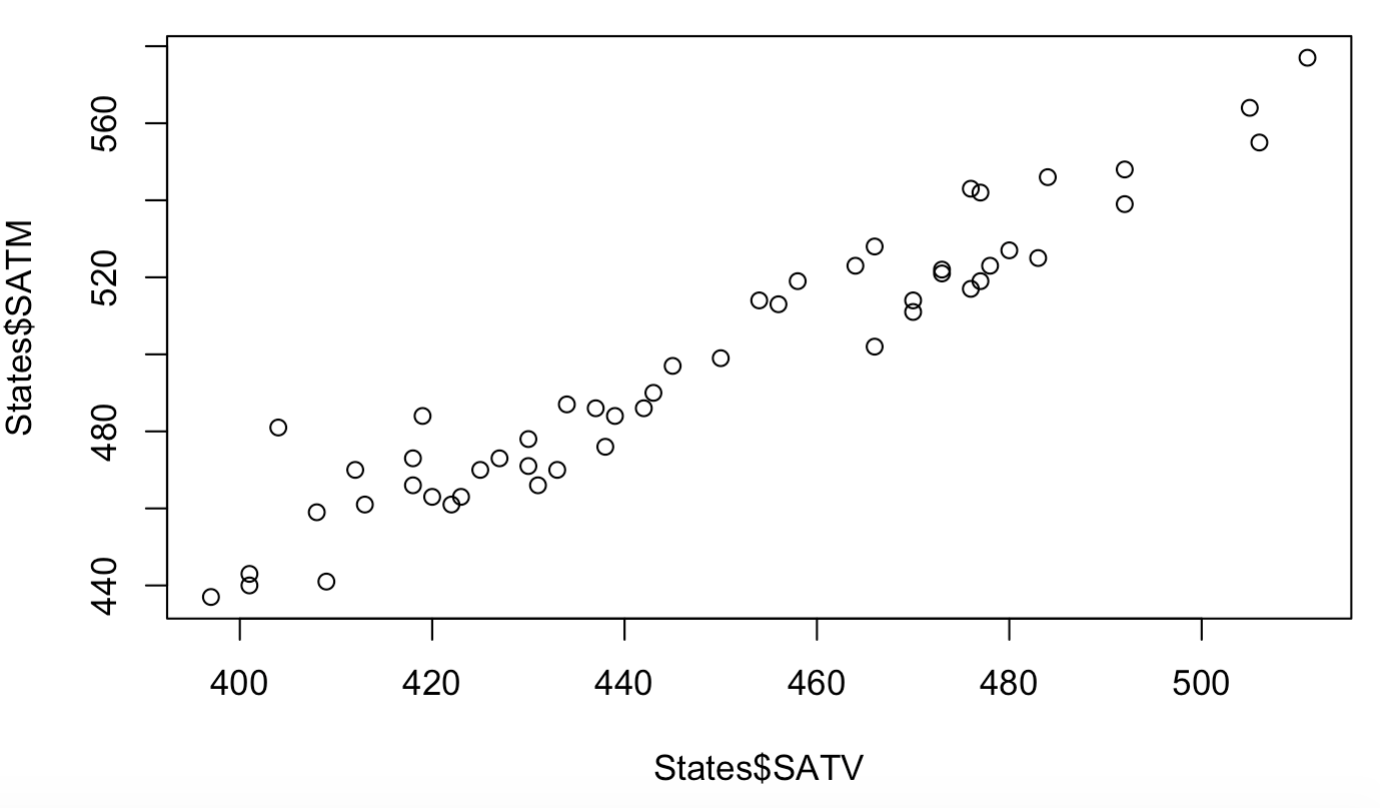
*> plot(States$SATV, States$SATM)*

Figure 9. Scatterplot

**Conclusion:** The data is more or less along a straight line, without a big variation. It has a high positive correlation.

**(c)  Compute Pearson’s and Spearman’s coefficient of correlation. Interpret and compare their values. Are their values consistent with the scatter plot?**

**Code:**

*> cor(States$SATV, States$SATM,method="spearman")*

***[1] 0.9481642***

*> cor(States$SATV, States$SATM,method="pearson")*

***[1] 0.9620359***

**Conclusion:** the relative strength of the linear relationship is positively strong, coefficients values are very close to 1.

**(d)  Add the marginal distributions to the scatter plot. For that purpose, use histogram and box plot.**

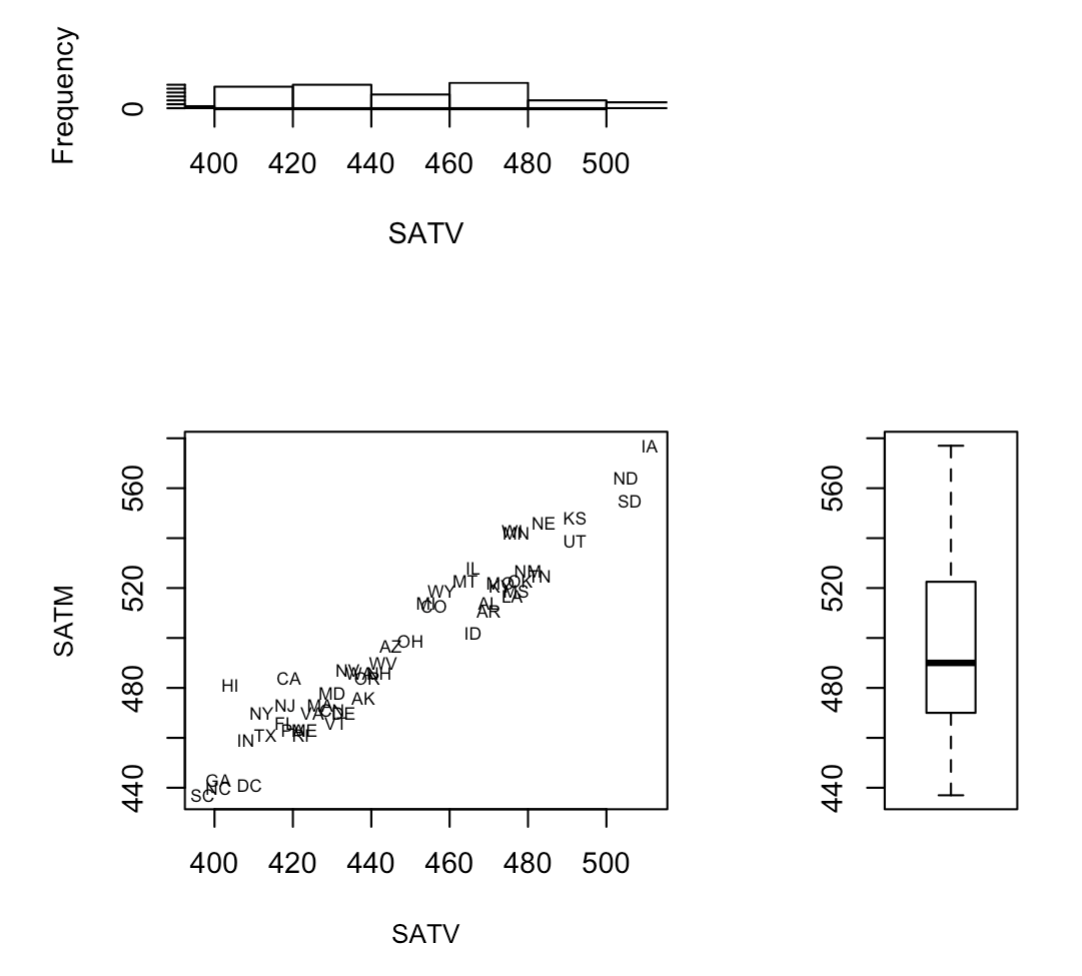
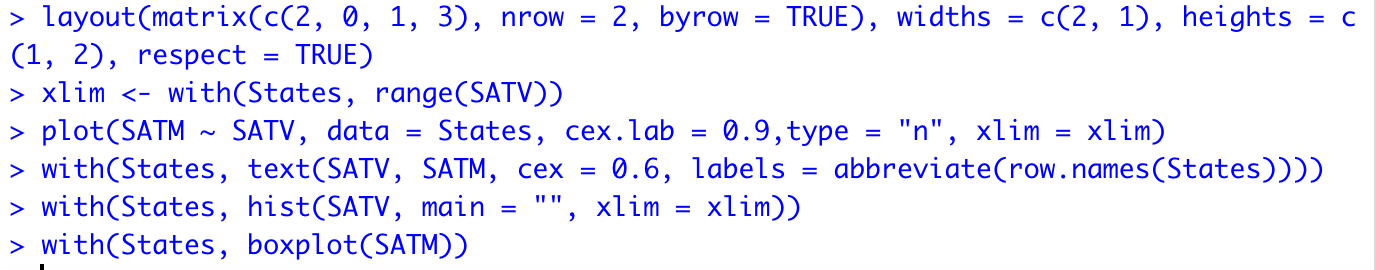
Code:

Figure 10. Scatterplot that shows the marginal distributions by histogram and boxplot.

**Conclusion:** According to the scatterplot there are no outliers in this data.

**(e)  Depict the bivariate box plot. Comment on the outliers. Remove the outliers, if any, and re-compute the Pearson correlation coefficient.**

***Code:***

*outcity <- match(lab <- c("CA"), rownames(States))*

*> x <- States[, c("SATV", "SATM")]*

*> bvbox(x, mtitle = «")*

*> text(States$SATV[outcity], States$SATM[outcity], labels = lab, cex = 0.7, pos = c(2, 2, 4, 2, 2))*

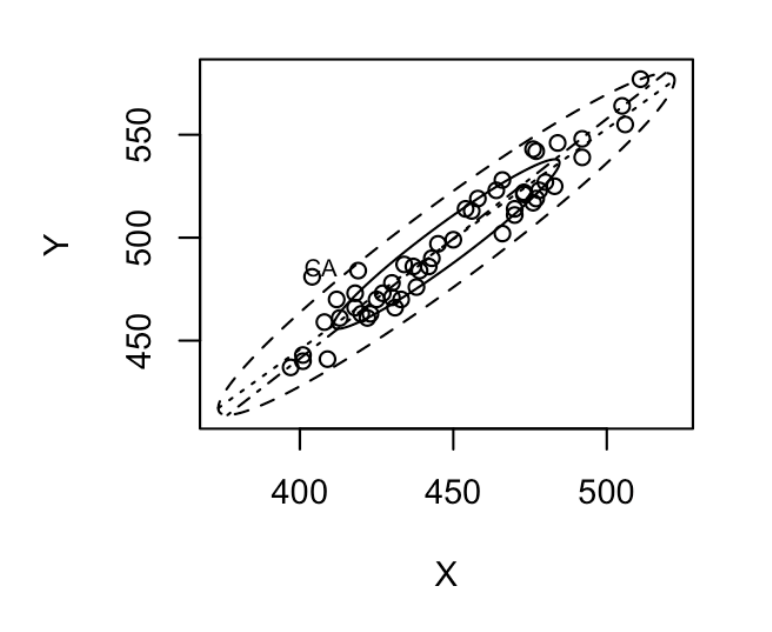
****

Figure 10. Scatterplot showing the bivariate boxplot of the data.

**Conclusion:** This type of graphic may be useful in indicating the distributional properties of the data and in identifying possible outliers. Here we see that dataset does not contain any real outliers. Only one value is above the line so it was commented.

**(f)  Create the convex hull. Remove the observations lying on the hull and re-compute the correlation coefficient.**

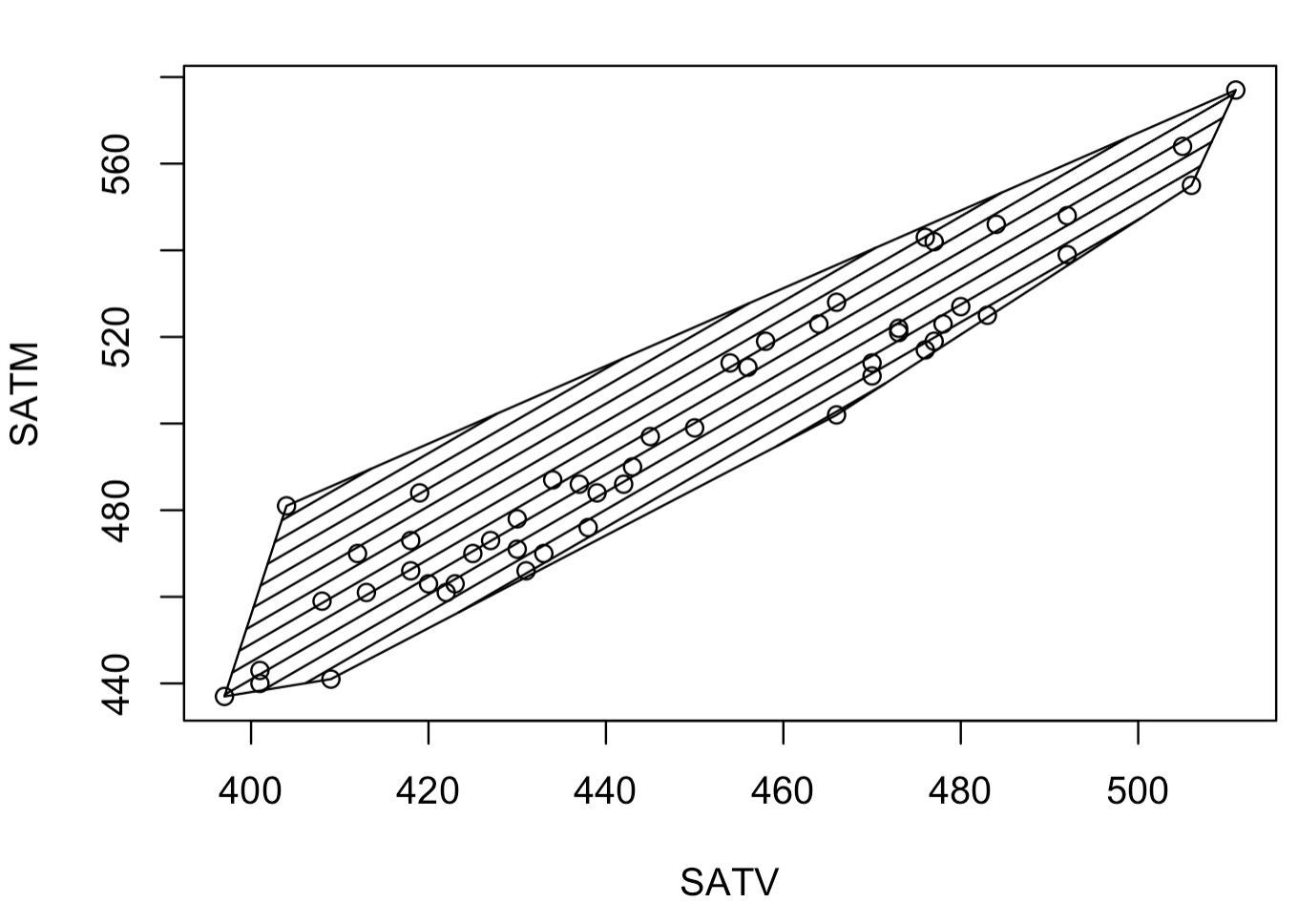
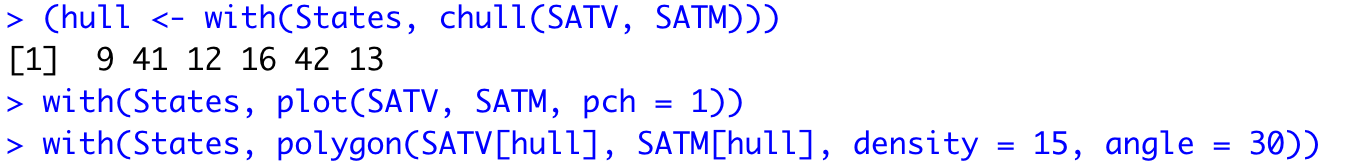
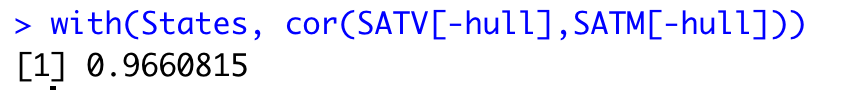
***Code:***

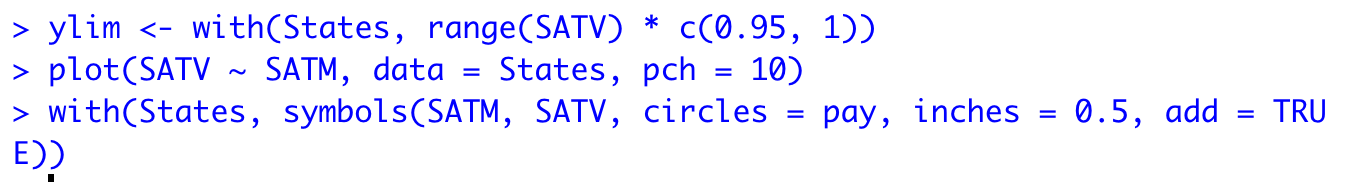
Figure 11. convex hull of the data

****

**Conclusion:** correlation coefficient increased when the points on the convex hull were deleted.

**3. Multivariate data:**

**(a)  Pick up a dataset which has three variables (from source 2 or 3) and create the bubble plot. Interpret the result.**

**Code:**

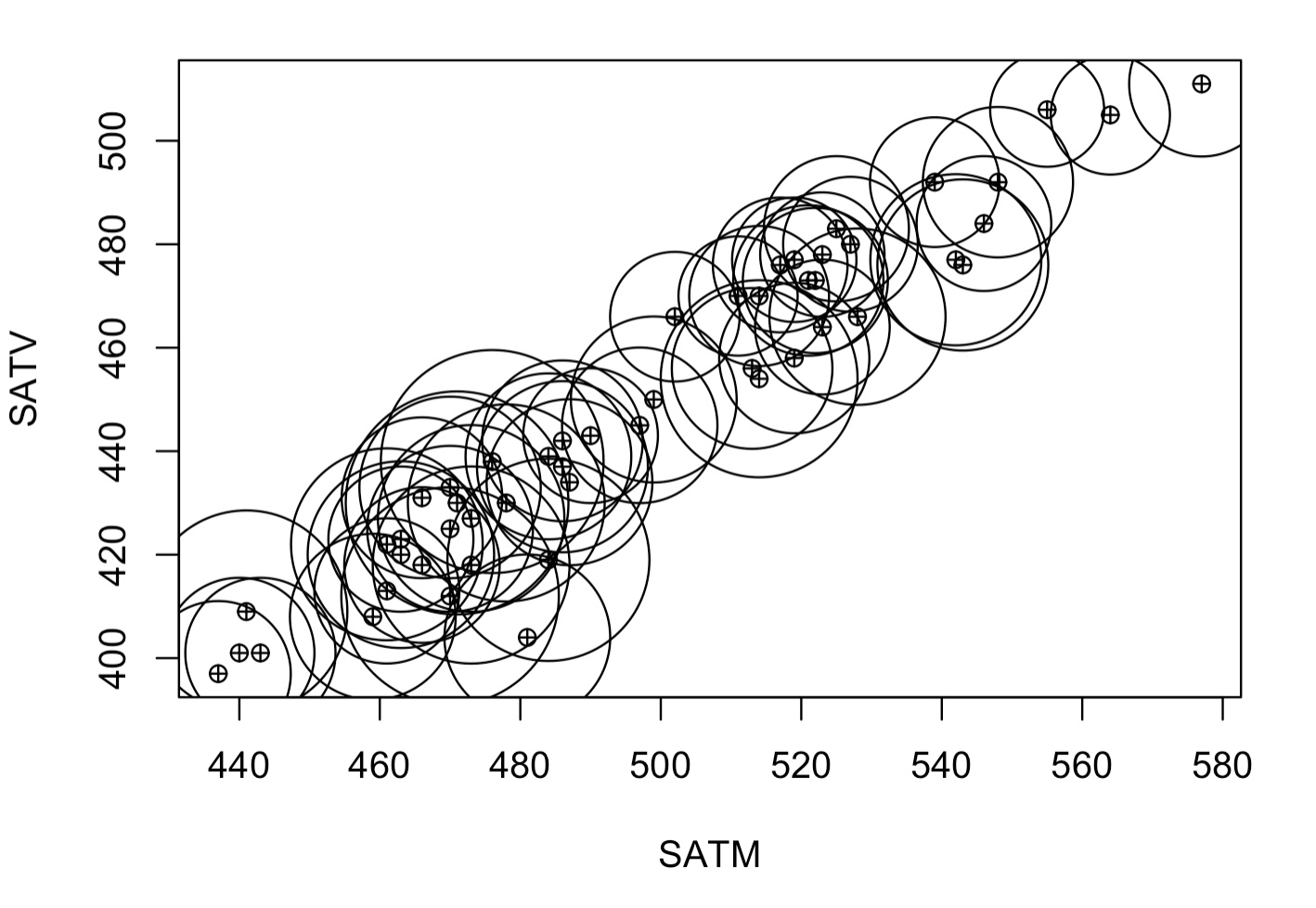
****

Figure 12: bubble plot on three variables: SATV, SATM, pay

**Conclusion:** bubble plot allows to display three variables; two are used to form the scatterplot itself, and then the values of the third variable are represented by circles with radii proportional to these values and centred on the appropriate point in the scatterplot

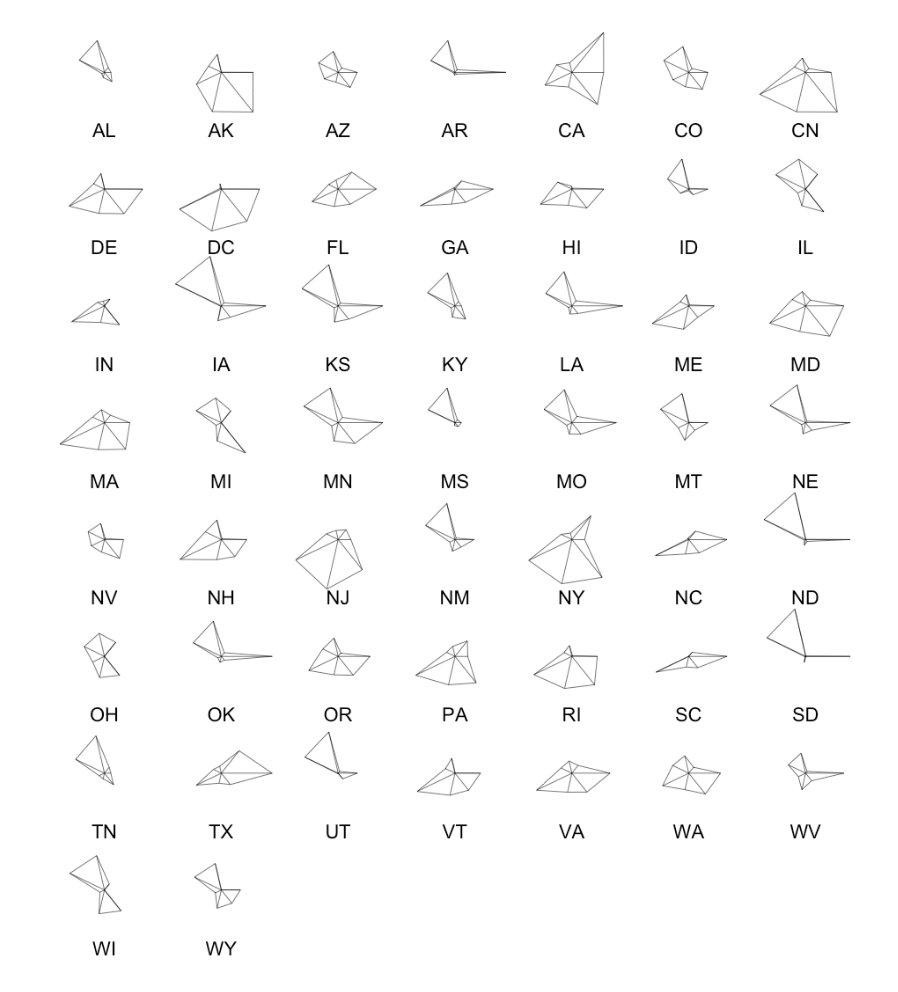
**(b)  Use data source 2 or 3. Create the glyph plot of all observations, Section 2.3. Do any stars look alike?**   


Figure 13: Star plot

**Conclusion:** there are several stars that looks quite the same: IA and KS have the closest star shapes.

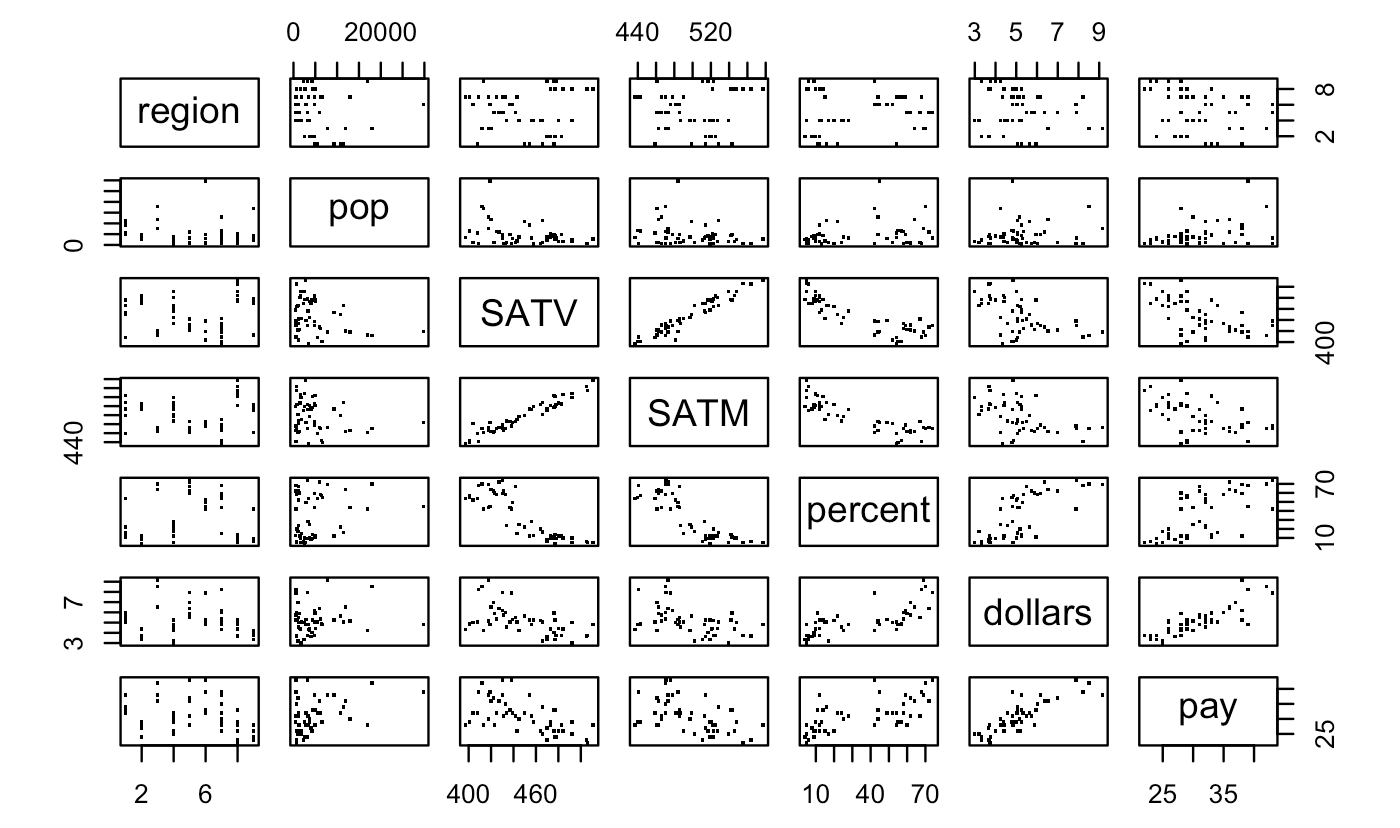
**(c)  Use data source 2 or 3. Create the scatter plot matrix and analyze it. See [2], Section 2.4.**   


Figure 14: Scatterplot matrix

**Conclusion:** SATV and SATM has the strongest linear relationship.