Khan Inan

1. For this specific dataset we can see that there are certain star players and some players who don't quite have the stats to match. Based on the data presented here certain players seem to stand out like Kobe Bryant and Lebron James that are top 3 in a lot of the categories. They must have had tremendous skill and team assistance to achieve these kinds of numbers and during their peak they have numbers greater than every other all-star player.

   In terms of successes and failures for some of the players, it becomes obvious that a player is falling off or their skill is degrading due to a variety of factors, such as age, an injury, a not so competent team, etc. All these factors contribute to a players numbers decreasing after they hit their peak, because not every players stats are going to increase forever from season to season, no matter how good they are. There are ways to predict a breakthrough season for a player however. If we can sees a players stats going up year by year consistently, it can be reasonable to predict that they will continue to go up. If their stats fluctuate, however, then it is more difficult

   Based on some of the quantile plot data below that is a mix of the players, we can see an interesting phenomenon. It is that, for some particular stats like rebounds and assists, no matter how good the player is in everything else, all the players still fall in the same range of values. This is really interesting because it may suggest some kind of soft limit to how high these stats can possibly get, possibly due to the rules and structure of the game of basketball itself, or maybe even more interestingly, the limits of the human body.

2.

```
> as.matrix.data.frame(train)
       V1      V2        V3        V4        V5        V6        V7        V8        V9        V10       V11       V12
[1,] "label"  "pixel0"  "pixel1"  "pixel2"  "pixel3"  "pixel4"  "pixel5"  "pixel6"  "pixel7"  "pixel8"  "pixel9"  "pixel10"
       V13     V14       V15       V16       V17       V18       V19       V20       V21       V22       V23
[1,] "pixel11" "pixel12" "pixel13" "pixel14" "pixel15" "pixel16" "pixel17" "pixel18" "pixel19" "pixel20" "pixel21"
       V24     V25       V26       V27       V28       V29       V30       V31       V32       V33       V34
[1,] "pixel22" "pixel23" "pixel24" "pixel25" "pixel26" "pixel27" "pixel28" "pixel29" "pixel30" "pixel31" "pixel32"
       V35     V36       V37       V38       V39       V40       V41       V42       V43       V44       V45
[1,] "pixel33" "pixel34" "pixel35" "pixel36" "pixel37" "pixel38" "pixel39" "pixel40" "pixel41" "pixel42" "pixel43"
       V46     V47       V48       V49       V50       V51       V52       V53       V54       V55       V56
[1,] "pixel44" "pixel45" "pixel46" "pixel47" "pixel48" "pixel49" "pixel50" "pixel51" "pixel52" "pixel53" "pixel54"
       V57     V58       V59       V60       V61       V62       V63       V64       V65       V66       V67
[1,] "pixel55" "pixel56" "pixel57" "pixel58" "pixel59" "pixel60" "pixel61" "pixel62" "pixel63" "pixel64" "pixel65"
       V68     V69       V70       V71       V72       V73       V74       V75       V76       V77       V78
[1,] "pixel66" "pixel67" "pixel68" "pixel69" "pixel70" "pixel71" "pixel72" "pixel73" "pixel74" "pixel75" "pixel76"
       V79     V80       V81       V82       V83       V84       V85       V86       V87       V88       V89
[1,] "pixel77" "pixel78" "pixel79" "pixel80" "pixel81" "pixel82" "pixel83" "pixel84" "pixel85" "pixel86" "pixel87"
       V90     V91       V92       V93       V94       V95       V96       V97       V98       V99       V100
```
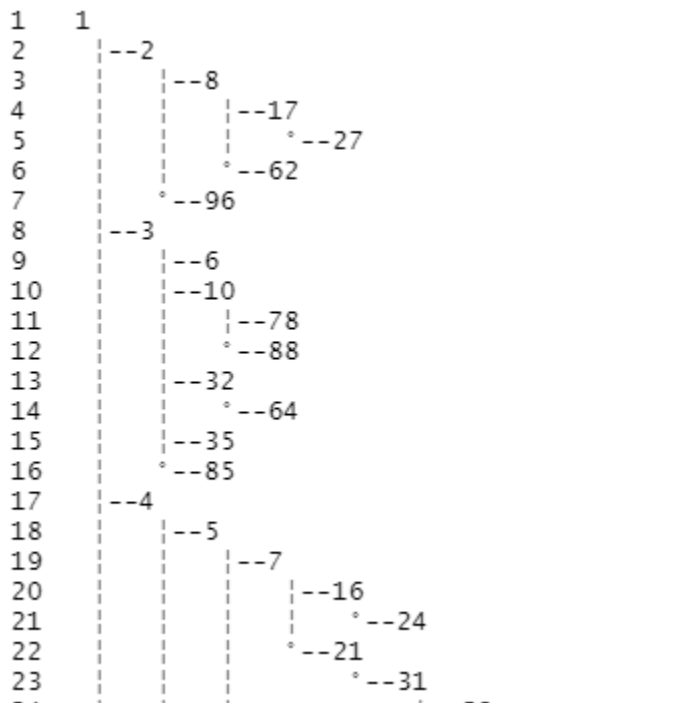
   Visualizing the data using a matrix we can see that the numbers align for some aspects of the pixels but not for some pixels in the middle or towards the end. In order to reproduce the digits reasonably well, I would say there should be some automation for pixel alignment. My final matrix is semi-aligned

The creation of a data tree was fairly tricky, given that there are many options

```
1    1
2     |--2
3     |   |--8
4     |   |   |--17
5     |   |   |    °--27
6     |   |   °--62
7     |   °--96
8     |--3
9     |   |--6
10    |   |--10
11    |   |   |--78
12    |   |   °--88
13    |   |--32
14    |   |   °--64
15    |   |--35
16    |   °--85
17    |--4
18    |   |--5
19    |   |   |--7
20    |   |   |   |--16
21    |   |   |   |   °--24
22    |   |   |   °--21
23    |   |   |        °--31
```

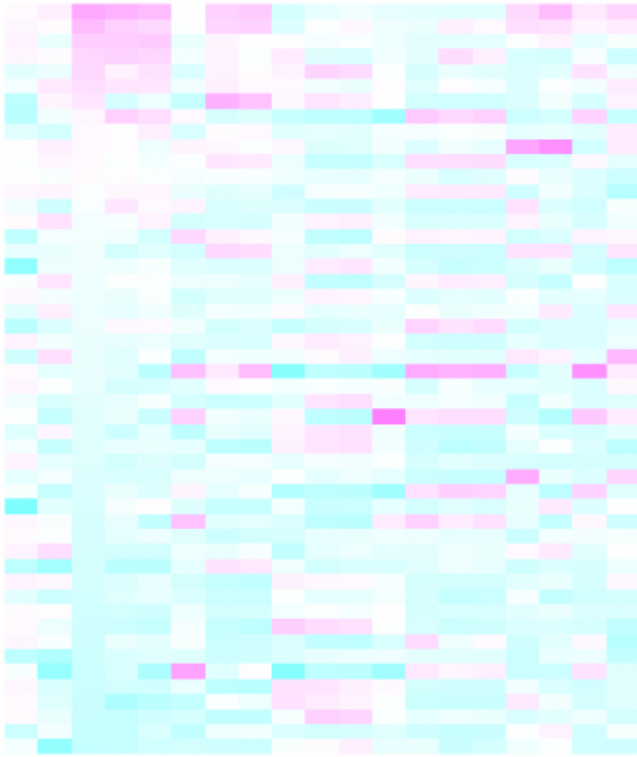The distance between the pixels is shown in the tree and as we can see it varies wildly. based on this and the geometry of the pixels/ PCA results, We can conclude that the geometry of the pixels and how far apart they are in the 2d space is inconsistent

For taking the average of all the data for each number and creating the dot product and identifying pixels. I obtained a fairly random dotchart

3.
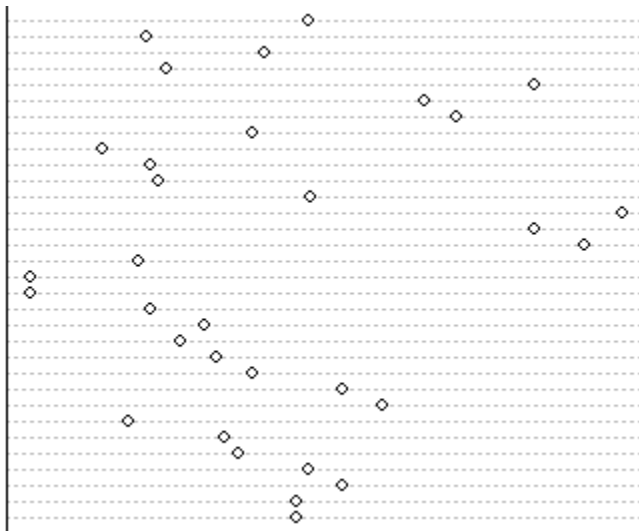
      Building a hierarchical tree for this mnemiopsis data, I obtained something that looked like this that continues for a fair bit of rows

```
 1   1
 2      |--2
 3      |--3
 4      |   |--4
 5      |   |   |--6
 6      |   |   |   |--7
 7      |   |   |   |    °--28
 8      |   |   |   |        °--101
 9      |   |   |   |--8
10      |   |   |   |    °--33
11      |   |   |   |        °--81
12      |   |   |   |--9
13      |   |   |   |    °--68
14      |   |   |   |        °--70
15      |   |   |       °--52
16      |   |   |--11
17      |   |   |   |--12
18      |   |   |   |    |--21
19      |   |   |   |    |--24
20      |   |   |   |    |    °--42
21      |   |   |   |    |        |--47
22      |   |   |   |    |        |    °--76
23      |   |   |   |    |        °--56
24      |   |   |   |    |            °--58
25      |   |   |   |    |--31
26      |   |   |   |    |--37
27      |   |   |   |    |    |--38
28      |   |   |   |    |    |    °--53
29      |   |   |   |    |    |        °--71
30      |   |   |   |    |    |            |--72
31      |   |   |   |    |    |            |--77
32      |   |   |   |    |    |                °--98
33      |   |   |   |    |    |--82
34      |   |   |   |    |    |--88
35      |   |   |   |    |    °--94
36      |   |   |   |    |        41
```

drawing a heatmap in R using the heatmap() expression and mnemiopsis data I get

Using deseq2 to analyze this data, the first thing that I found was that the most significantly changing genes within this data are the genes ML14112a, ML26798a, ML06971a, ML076034a, ML22162a, ML017310a, ML205622a and ML20395a

The genes that are the most highly expressed are represented as the darker, orange and red colors within the heatmap. They are the housekeeping genes because a lot of other genes and bodily functions depend on those genes to be able to exist and work properly

These results I would say are fairly consistent with the mid term project because, at that time I was unable to see the entire context of the data. Meaning that I knew there were genes more highly expressed than others, but I didn't see the extent to which there are genes that are less expressed (green)

In terms of consistency, what I can say about this data is that the less expressed genes outnumber the highly expressed genes by a factor of something like 15 to 1. This result is very interesting because it puts into perspective how just a few important genes create the blueprint and instructions for the vast majority of the genes in the human body. This information can be extrapolated to determine something like a tree of which gene depends on what. This would give a lot of insight into genetic and chromosomal deformities, and where they originate from at birth or in the body

The most interesting pathway/gene that is responding in this study, is the darkest cell within the heatmap which happens to be ML46651a. This just so happens to be part of an epithelial cell, which are cells that line every part of the human body, from skin, to blood vessels, to organs. After looking at what this gene does for these very important and numerous epithelial cells, it makes sense why it would be the darkest on this heatmap