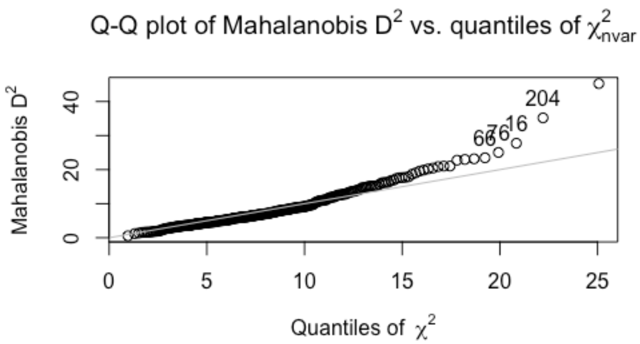
Holly Ghaemi
1092917

Multivariate Homework 2

Question 1

1. The outliers in this data set are the data points which deviate from the QQ line.



D^2 statistic is a popular metric called the Mahalonobis Distance and it describes the distance between an observation $\boldsymbol{Y_i}$and the multivariate mean $\overline{\boldsymbol{Y}}$

$$D_i^2 = (\boldsymbol{Y_i} - \overline{\boldsymbol{Y}})'\boldsymbol{S^{-1}}(\boldsymbol{Y_i} - \overline{\boldsymbol{Y}})'$$

The above graph shows the scalar D_i^2 values against chi-squared quantiles. Since the components of $\boldsymbol{Y_i}$ are correlated and have different variances, a simple Euclidean distance would not be appropriate. T

2. Principal component analysis is the identification of linear combination of variables that provides the maximum variability. The first principal component has the greatest variability. The second component has the maximum variability among all linear combinations that are orthogonal to the first. The third principal component is orthogonal to both the first and second and so on for further component analysis. PCA reduces a large number of multivariate variables into a relatively small number of linear combinations that can be used to account for much of the variability in the data. Variables with the greatest variance will typically dominate the analysis.

The weights, w_{ij}, are also called loadings because they explain how much each of the original observations, x_i contributes to each of the principal components. They are chosen so that the y_i have the largest possible variances and are mutually uncorrelated.

Covariance is defined with the following equation: $\widehat{Cov}(y_i, y_j) = e_i'\boldsymbol{S}e_j = \lambda_j e_i' e_j$. It is important to note that covariance is not divided by standard deviation and therefore it is not scaled. This will be further addressed in the following section.

Holly Ghaemi
1092917

Covariance without Outliers

```
> affectPC.nooutliers$loadings

Loadings:
   Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
X1  0.553        -0.365 -0.533 -0.115  0.292  0.417       
X2         0.327               -0.808 -0.195         0.430
X3  0.490        -0.509  0.220  0.144 -0.419 -0.481 -0.137
X4         0.249         0.332 -0.359  0.512 -0.101 -0.646
X5  0.512         0.674 -0.180         0.189 -0.461       
X6         0.694  0.234 -0.273  0.148 -0.417  0.251 -0.360
X7  0.436         0.241  0.651        -0.125  0.544  0.114
X8         0.582 -0.158  0.158  0.398  0.462         0.484

               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000
Proportion Var  0.125  0.125  0.125  0.125  0.125  0.125  0.125  0.125
Cumulative Var  0.125  0.250  0.375  0.500  0.625  0.750  0.875  1.000
```

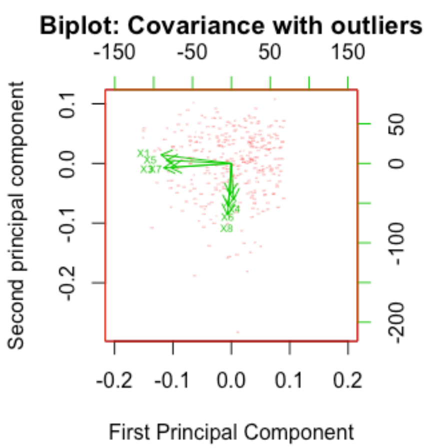
Covariance with outliers

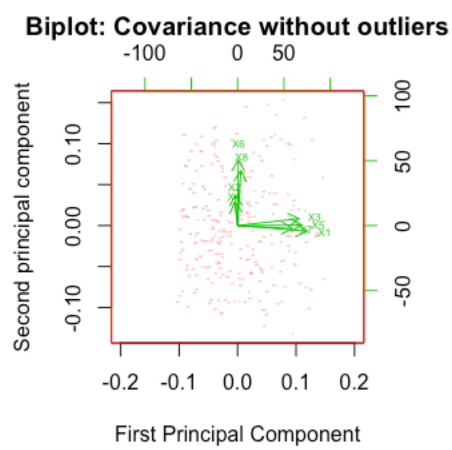
```
> affectPC$loadings

Loadings:
   Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
X1 -0.532  0.104 -0.450 -0.290  0.310  0.336  0.446 -0.112
X2        -0.371 -0.221  0.366  0.601 -0.295         0.480
X3 -0.512        -0.443        -0.332 -0.386 -0.523       
X4        -0.435         0.500         0.428 -0.161 -0.581
X5 -0.492         0.593         0.327  0.259 -0.449  0.142
X6        -0.516  0.237 -0.491  0.175 -0.458  0.131 -0.423
X7 -0.460         0.369  0.464 -0.324 -0.217  0.531       
X8        -0.624        -0.255 -0.429  0.378         0.464

               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000
Proportion Var  0.125  0.125  0.125  0.125  0.125  0.125  0.125  0.125
Cumulative Var  0.125  0.250  0.375  0.500  0.625  0.750  0.875  1.000
```

Comparison: The first component in the data without outliers is primarily made up of X1 and X5 where X1 = EA1 and X5 = EA2. The second component of this data is made up of X6 = TA2 and X8 = NA2. The first component in the data with outliers is made up of mostly X1 = EA1 and X3 = PA1. The second component is mostly made up of X6 = TA2 and X8 = NA2. The first components in each are made up of varying variables and they are opposite in magnitude. In the second component, the largest components are the same but they are also opposite in magnitude. It seems that outlier removal has an effect on the magnitude of the loadings. We can see opposite magnitudes from the biplots below as well. It is apparent that PCA is very sensitive to outliers.





Biplots are a graphical method to help interpret the first two components and the first two principal components are shown using arrows that indicate their directions. We can see from these biplots that the directions of the components vary with and without outliers.

3. *Comparison:* The first component without outliers is primarily made up of X1, X3 and X7. The first component with outliers is primarily made up of X1, X3 and X7. Once again, we can see that the components are made up of the same variables but they are opposite in magnitude. The second component without outliers is primarily made up of X8 and X4. The second component with outliers is primarily made up of X6 and X8. The components here are different but they are all of opposite magnitudes yet again. Outliers not only

affect which variables make up the greatest proportion in each component but also change the magnitudes.

Correlation without Outliers

```
> affectPC.corr$loadings

Loadings:
   Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
X1 -0.496               -0.412  0.311  0.463         0.505
X2        -0.469  0.523 -0.490 -0.231 -0.230  0.399
X3 -0.511                       0.470 -0.397 -0.246 -0.534
X4        -0.530  0.430  0.409         0.375 -0.474
X5 -0.489        -0.156        -0.543  0.427  0.208 -0.459
X6        -0.463 -0.583 -0.343 -0.268 -0.180 -0.458  0.108
X7 -0.496                0.477 -0.274 -0.454         0.485
X8        -0.524 -0.395  0.265  0.434  0.115  0.542

               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000
Proportion Var  0.125  0.125  0.125  0.125  0.125  0.125  0.125  0.125
Cumulative Var  0.125  0.250  0.375  0.500  0.625  0.750  0.875  1.000
```

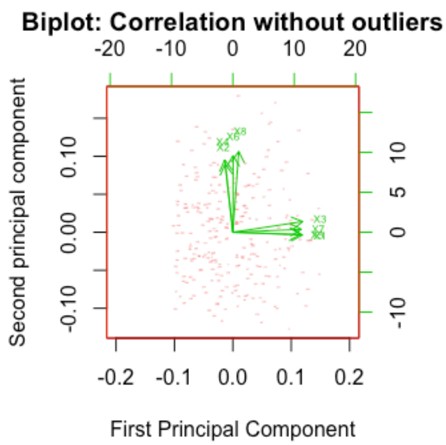
Correlation with Outliers

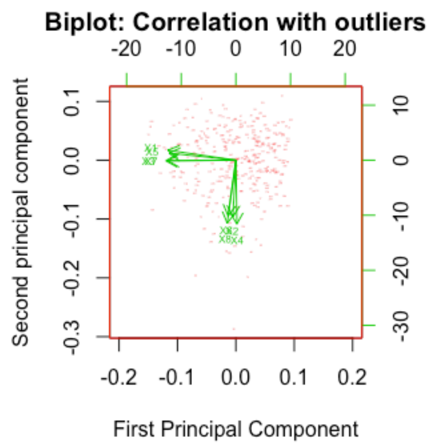
```
> affectPC.corr.noout$loadings

Loadings:
   Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
X1  0.504               -0.131 -0.479  0.496         0.503
X2         0.455 -0.469 -0.592 -0.202 -0.276  0.318
X3  0.505                0.215 -0.428 -0.337 -0.343 -0.523
X4         0.483 -0.542  0.362  0.249  0.425 -0.305
X5  0.488         0.116 -0.302  0.472  0.364  0.278 -0.474
X6         0.512  0.539 -0.354  0.127        -0.537  0.124
X7  0.495        -0.159  0.150  0.477 -0.499         0.480
X8         0.540  0.388  0.465 -0.138         0.567

               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000
Proportion Var  0.125  0.125  0.125  0.125  0.125  0.125  0.125  0.125
Cumulative Var  0.125  0.250  0.375  0.500  0.625  0.750  0.875  1.000
```

The varying magnitudes can also be seen in the following biplots:





4. The difference between the correlation PCA and Covariance PCA is as follows: the components are primarily made up of different variables and the magnitudes are different.

   Correlation is a better approach to use when deriving principal components because when the standard deviations of all the vectors are very different, it is important to represent them in the same scale.

   The following are the standard deviations of each vector in the data set with outliers and without outliers respectively.

Holly Ghaemi
1092917

```
> sapply(affect, sd)
      X1       X2       X3       X4       X5       X6       X7       X8
7.112951 4.420333 6.788680 4.308987 6.851012 4.883519 6.425724 5.191015
> sapply(affect.outliersremoved, sd)
      X1       X2       X3       X4       X5       X6       X7       X8
6.466533 3.516262 5.877289 2.890574 6.287399 4.383476 5.455640 3.929371
```

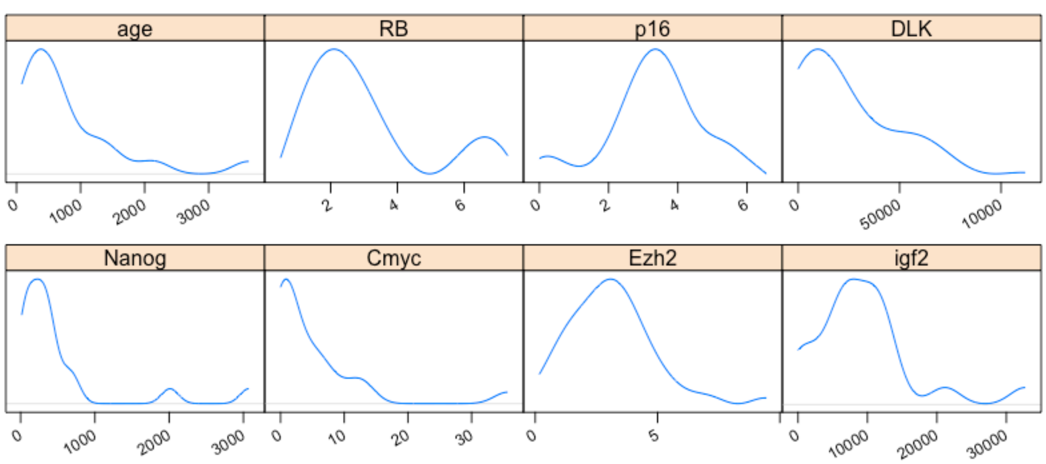
The standard deviations vary from with to without outliers as well as across the variables. The varying standard deviations affect PCA if the units are not scaled and the standard deviations are different.

When we work with the correlation matrix, all the units have been divided by their standard deviations unlike the covariance matrix. Therefore, they have been scaled. It is appropriate to use the covariance matrix when the units of the data are all the same, however it is a general rule to use the correlation matrix for PCA. As the data set 'affect' is a collection of pretest data using 5 scales from the Eysenck Personality Inventory, it is more appropriate to use a correlation matrix when conducting PCA.

Question 2

1. The following plots were used to detect outliers in the hemangioma data from table 8.2:

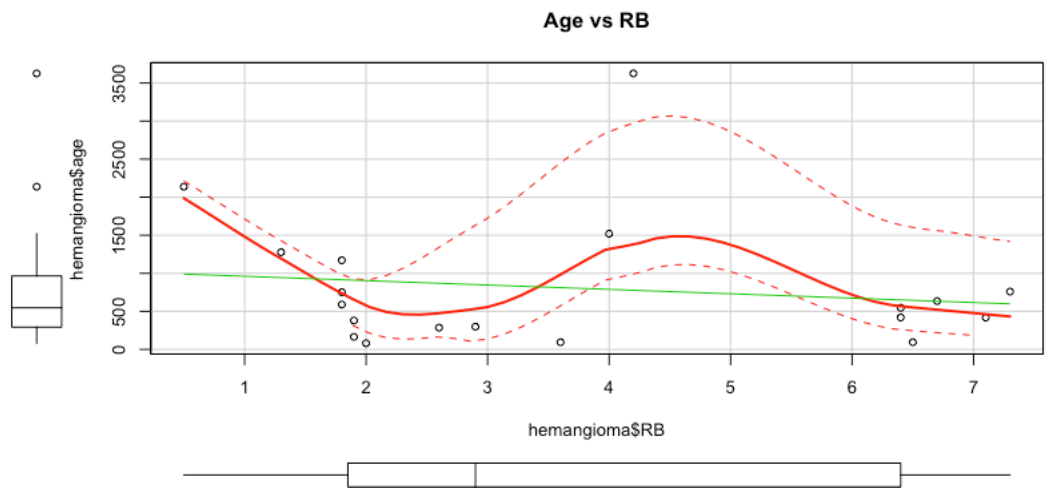
Marginal Plot

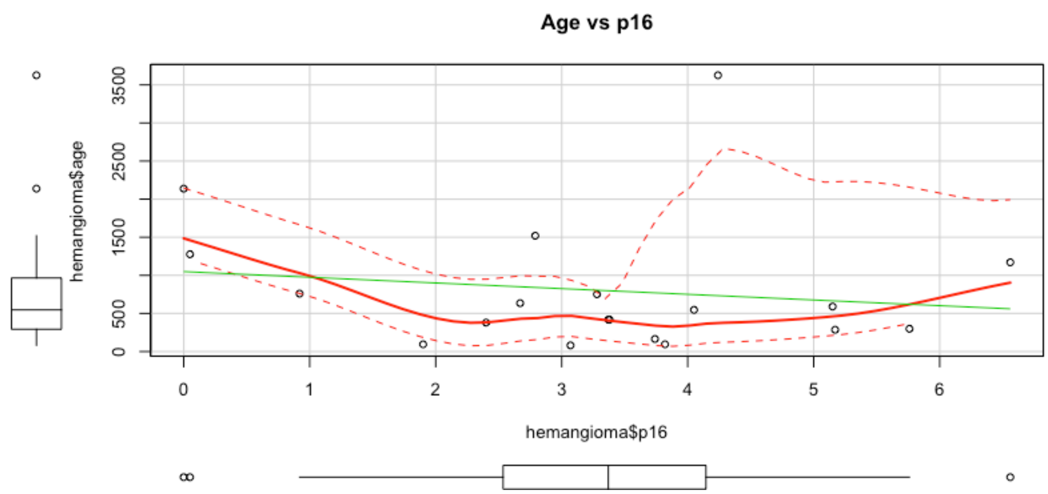


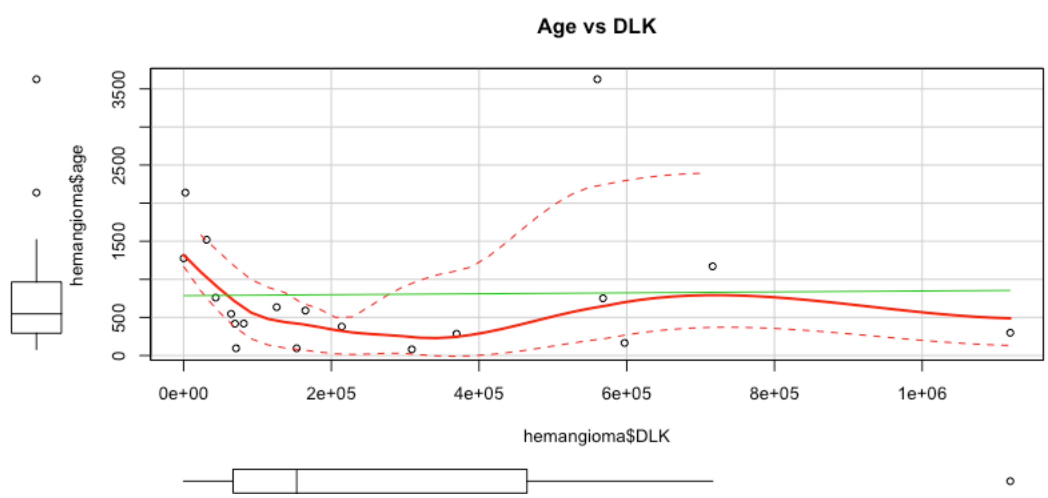
From the marginal plot we can see that age, cmyc, nanog, igf2, DLK. Ezh2 are right skewed. RB is bimodal. The only variable that shows the most normality is p16 although this case is also slight bimodal and left-skewed.

Scatterplots

Holly Ghaemi
1092917

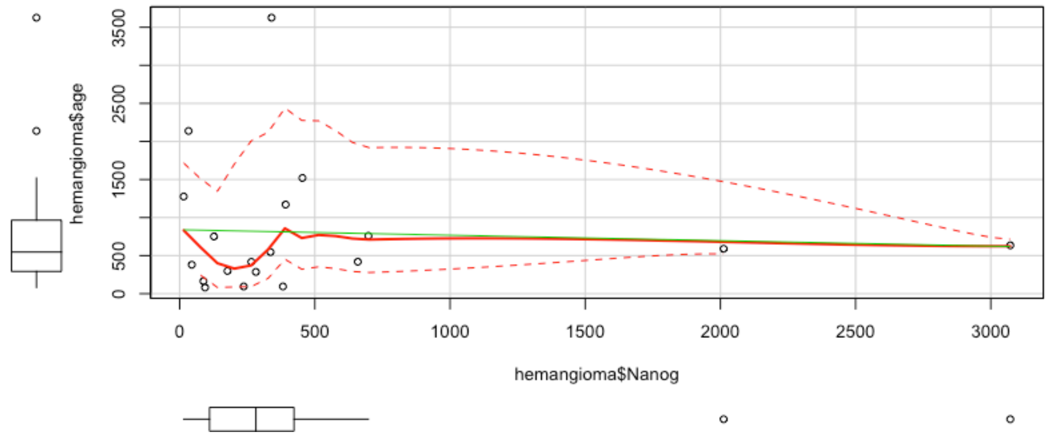




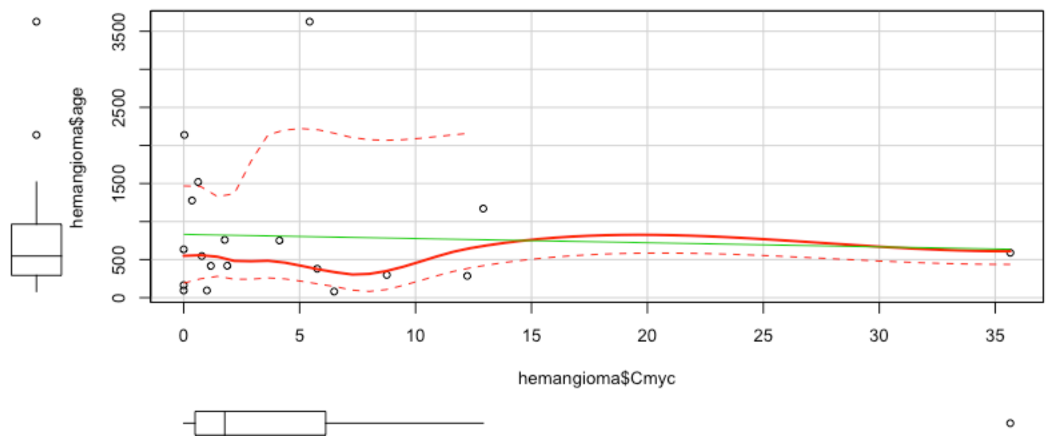


Holly Ghaemi
1092917

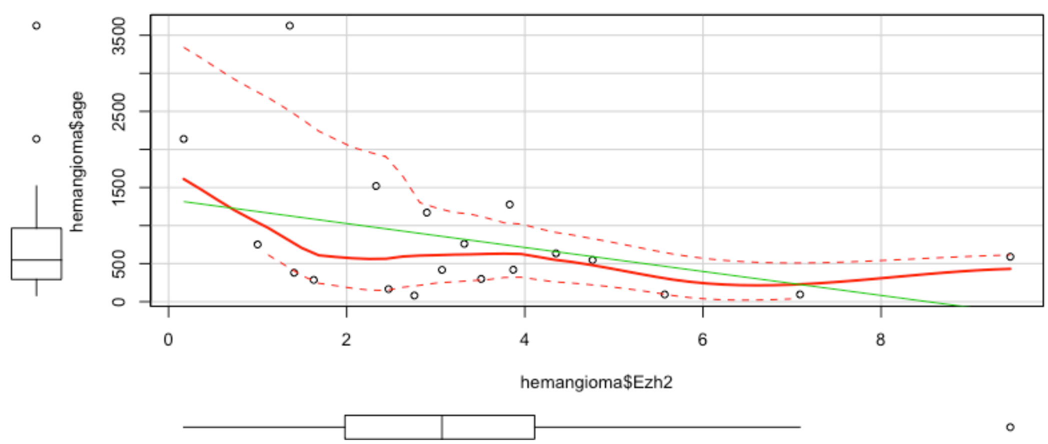
Age vs Nanog



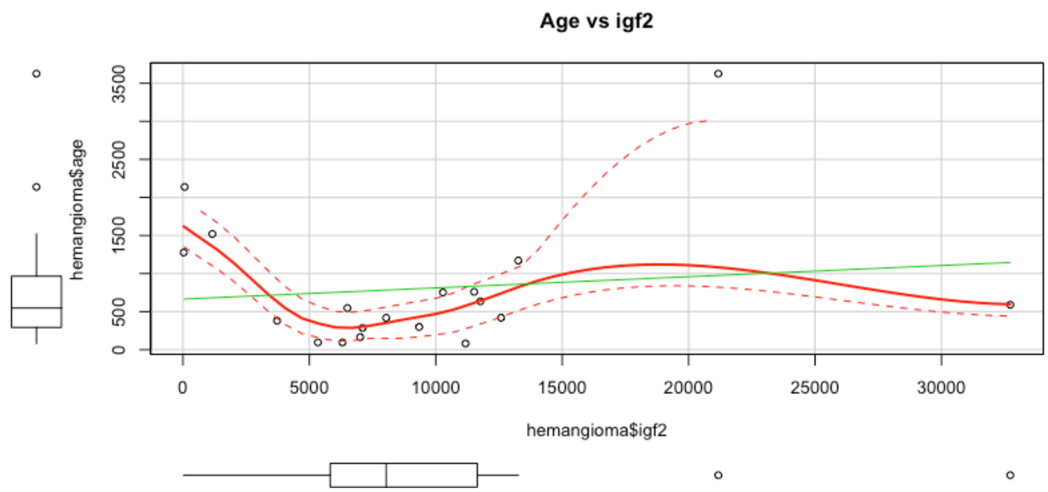
Age vs Cmyc



Age vs Ezh2



Holly Ghaemi
1092917



From the boxplots of age, we can see that there are two upper outliers. All of the variables besides RB have outliers as shown in their boxplots.

2. Factor analysis is a method for identifying the groups of variables (or *factors*) whose actions appear to work in parallel. Within a single factor, several measured variables within every individual are highly correlated whether positively or negatively. Other variables may act independently of the others. Factor analysis aims to identify and interpret these groups of factors and we must first begin with estimating the appropriate number of factors needed to model the data.

The factor analysis on the hemangioma data with the outliers included allows us to display the following table:

No. of Factors	p-value
1	0.0073
2	0.0622
3	0.33
4	0.519

The appropriate number of factors is anywhere from 2-4 according to the p-values as we fail to reject the null hypothesis that the no. of factors is significant. However we must work at interpreting the output carefully because we cannot rely entirely on p-values.

Holly Ghaemi
1092917

Factor 1

```
Loadings:
      Factor1
age
RB
p16    0.575
DLK    0.260
Nanog  0.508
Cmyc   0.826
Ezh2   0.508
igf2   0.938

               Factor1
SS loadings      2.478
Proportion Var   0.310
```

Factor 2

```
Loadings:
      Factor1 Factor2
age
RB     0.139  -0.366
p16    0.369   0.731
DLK            0.961
Nanog  0.632  -0.159
Cmyc   0.755   0.341
Ezh2   0.671  -0.181
igf2   0.848   0.325

               Factor1 Factor2
SS loadings      2.305   1.871
Proportion Var   0.288   0.234
Cumulative Var   0.288   0.522
```

Factor 3

```
Loadings:
      Factor1 Factor2 Factor3
age                   -0.168
RB            -0.170   0.964
p16    0.381   0.747
DLK            0.971  -0.211
Nanog  0.475           0.389
Cmyc   0.930   0.280  -0.227
Ezh2   0.614  -0.115   0.343
igf2   0.767   0.351   0.187

               Factor1 Factor2 Factor3
SS loadings      2.225   1.751   1.364
Proportion Var   0.278   0.219   0.170
Cumulative Var   0.278   0.497   0.668
```

Factor 4

```
Loadings:
      Factor1 Factor2 Factor3 Factor4
age                            0.899
RB            -0.171   0.862  -0.114
p16    0.389   0.728          -0.138
DLK            0.969  -0.207
Nanog  0.589           0.337
Cmyc   0.850   0.236  -0.458
Ezh2   0.645  -0.112   0.205  -0.422
igf2   0.864   0.327           0.172

               Factor1 Factor2 Factor3 Factor4
SS loadings        2.4   1.682   1.166   1.061
Proportion Var     0.3   0.210   0.146   0.133
Cumulative Var     0.3   0.510   0.656   0.789
```

Taking a closer look for a better interpretation, we notice that Factor three has loadings that are all small and therefore this can be interpreted as noise. Despite the pvalue of 3 factors being much larger than that of 2 factors, we can see from the loadings that 2 factors are all that is necessary here.

3. The factor analysis on the data without the outliers included allows us to display the following table:

No. of Factors	p-value
1	0.00583
2	0.586
3	0.738
4	0.523

From the p-values we would assume that 3 factors would provide the best analysis as it is the largest p-value.

Holly Ghaemi
1092917

1 factor

Loadings:

	Factor1
age	-0.355
RB	0.998
p16	
DLK	-0.469
Nanog	0.763
Cmyc	-0.320
Ezh2	0.603
igf2	0.319

	Factor1
SS loadings	2.490
Proportion Var	0.311

2 factors

Loadings:

	Factor1	Factor2
age	-0.390	-0.409
RB	-0.130	0.989
p16	0.820	0.125
DLK	0.873	-0.361
Nanog	0.116	0.785
Cmyc	0.690	-0.232
Ezh2	-0.110	0.593
igf2	0.743	0.420

	Factor1	Factor2
SS loadings	2.657	2.490
Proportion Var	0.332	0.311
Cumulative Var	0.332	0.643

3 factors

Loadings:

	Factor1	Factor2	Factor3
age	-0.292		-0.894
RB	-0.214	0.816	0.436
p16	0.787		0.191
DLK	0.891	-0.324	
Nanog	0.118	0.960	
Cmyc	0.726		-0.132
Ezh2	-0.196	0.393	0.572
igf2	0.719	0.423	0.262

	Factor1	Factor2	Factor3
SS loadings	2.641	2.044	1.440
Proportion Var	0.330	0.255	0.180
Cumulative Var	0.330	0.586	0.766

4 factors

Loadings:

	Factor1	Factor2	Factor3	Factor4
age	-0.213		-0.953	-0.199
RB	-0.220	0.885	0.378	
p16	0.961	0.171	0.204	
DLK	0.784	-0.337		0.334
Nanog		0.873		0.175
Cmyc	0.719	-0.139		0.211
Ezh2	-0.148	0.419	0.542	-0.131
igf2	0.515	0.395	0.207	0.729

	Factor1	Factor2	Factor3	Factor4
SS loadings	2.437	2.042	1.436	0.776
Proportion Var	0.305	0.255	0.180	0.097
Cumulative Var	0.305	0.560	0.739	0.836

Taking a closer look at the loadings, we can see that factor three with the outliers taken out has a few larger loadings than when the outliers were not removed. Factor four has mostly small loadings and therefore we can attribute them to noise. Therefore, we can say that three factors is the best for analysis when we remove the outliers.

4. Between 2 and 3, the number of factors required goes from 2 to 3 when we exclude the outliers because there are larger loadings for 3 factors without outliers. The magnitudes also differ and for two factors there are more zero values when the outliers are removed.
5. Historically it has been well known that factor analysis can yield misleading conclusions. Small changes in the data values can change the analysis greatly. For this reason we see such varying differences in factor analysis with and without outliers as these values affect the value of the loadings, the magnitude of the loadings, the direction of the loadings and the number of factors that would be a best fit for the data.