PCA on **all** standardised data

20% sample of complete cleaned data set, then standardised (N=300800)

Variables = ABV, Appearance, Aroma, Palate, Taste, Rating

PCA Summary:

Importance of components:

PC1 PC2 PC3 PC4 PC5 PC6

Standard deviation 1.9143 0.9582 0.7434 0.62751 0.53334 0.43193

Proportion of Variance 0.6107 0.1530 0.0921 0.06563 0.04741 0.03109

Cumulative Proportion 0.6107 0.7638 0.8559 0.92150 0.96891 1.00000

Rotations (Principle Components):

PC1 PC2 PC3 PC4 PC5 PC6

ABV 0.2143872 -0.93091418 -0.190328520 0.13883045 -0.1542208 0.09029961

reviewAppearance 0.3840500 -0.04621287 0.892734695 0.13653731 -0.1838983 -0.03055637

reviewAroma 0.4366943 -0.05016525 0.002717715 -0.82203349 0.2918812 0.21410388

reviewPalate 0.4483923 0.08873735 -0.114895348 0.50464691 0.7205871 0.06288652

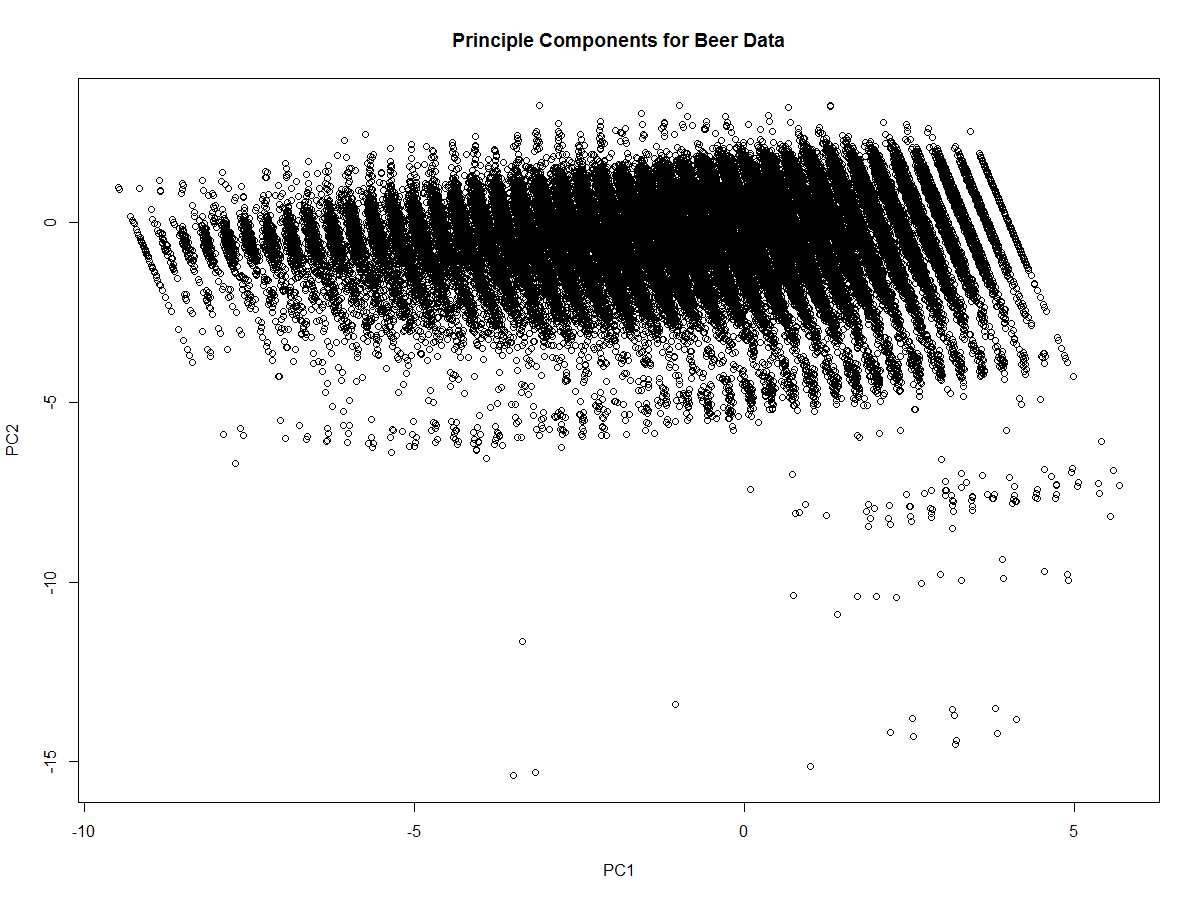
reviewTaste 0.4707906 0.11853162 -0.266027708 -0.05171331 -0.2444208 -0.79443054

rating 0.4394796 0.32683719 -0.287785945 0.17029926 -0.5274612 0.55677394

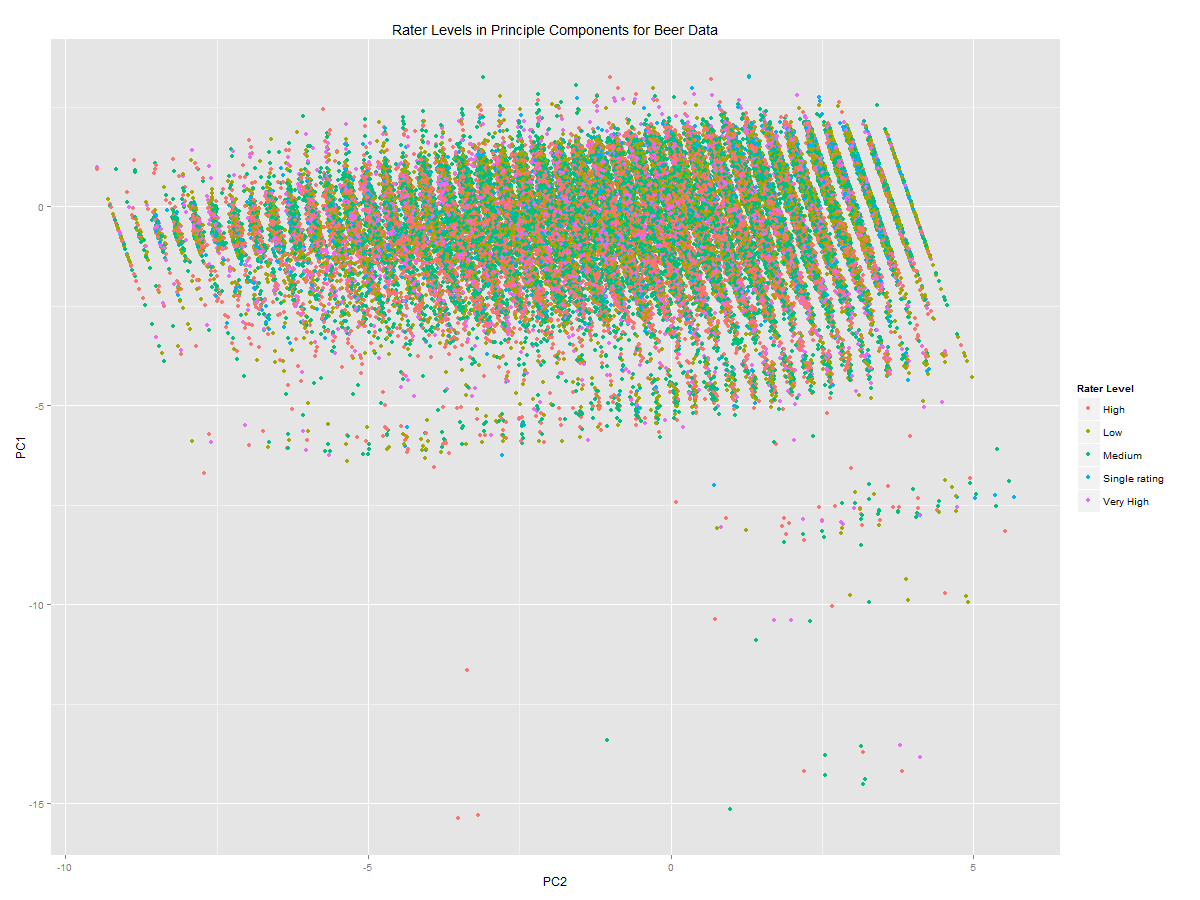
Mardia criterion for PCA1: 0.32955 [fulfilled by Appearance, Aroma, Palate, Taste and Rating]  
Mardia criterion for PCA2: 0.228786 [fulfilled by ABV and Rating]

Kaiser criterion suggests using only PC1 (as sd is > 1) however PC2 is very close to the cut off value

20% sample data plotted against PC1 and PC2. Slight negative correlation between PCs.



The frequency of each profile name was used to separate raters into bands:  
Very High raters >= 300 ratings  
High raters >=100 and <300 ratings  
Medium raters >=20 and <100 ratings  
Low raters >1 and <20 ratings  
Single raters = 1 rating  
Note: These bands were made up from the 20% data set, perhaps these raters have rated more in the complete data set!



The plot shows no discernable pattern – there are no obvious differences between higher and lower raters.  
\* Need to go back and swap axis names around!

PCA without Rating

20% sample of complete cleaned data set, then standardised

Variables = ABV, Appearance, Aroma, Palate, Taste

PCA Summary:

Importance of components:

PC1 PC2 PC3 PC4 PC5

Standard deviation 1.7411 0.9135 0.7162 0.62097 0.4853

Proportion of Variance 0.6063 0.1669 0.1026 0.07712 0.0471

Cumulative Proportion 0.6063 0.7732 0.8758 0.95290 1.0000

Rotations (Principle Components):

PC1 PC2 PC3 PC4 PC5

ABV 0.2771594 -0.95729106 0.006240666 -0.07466042 0.03410833

reviewAppearance 0.4384625 0.13816837 0.881281495 -0.02111960 0.10750367

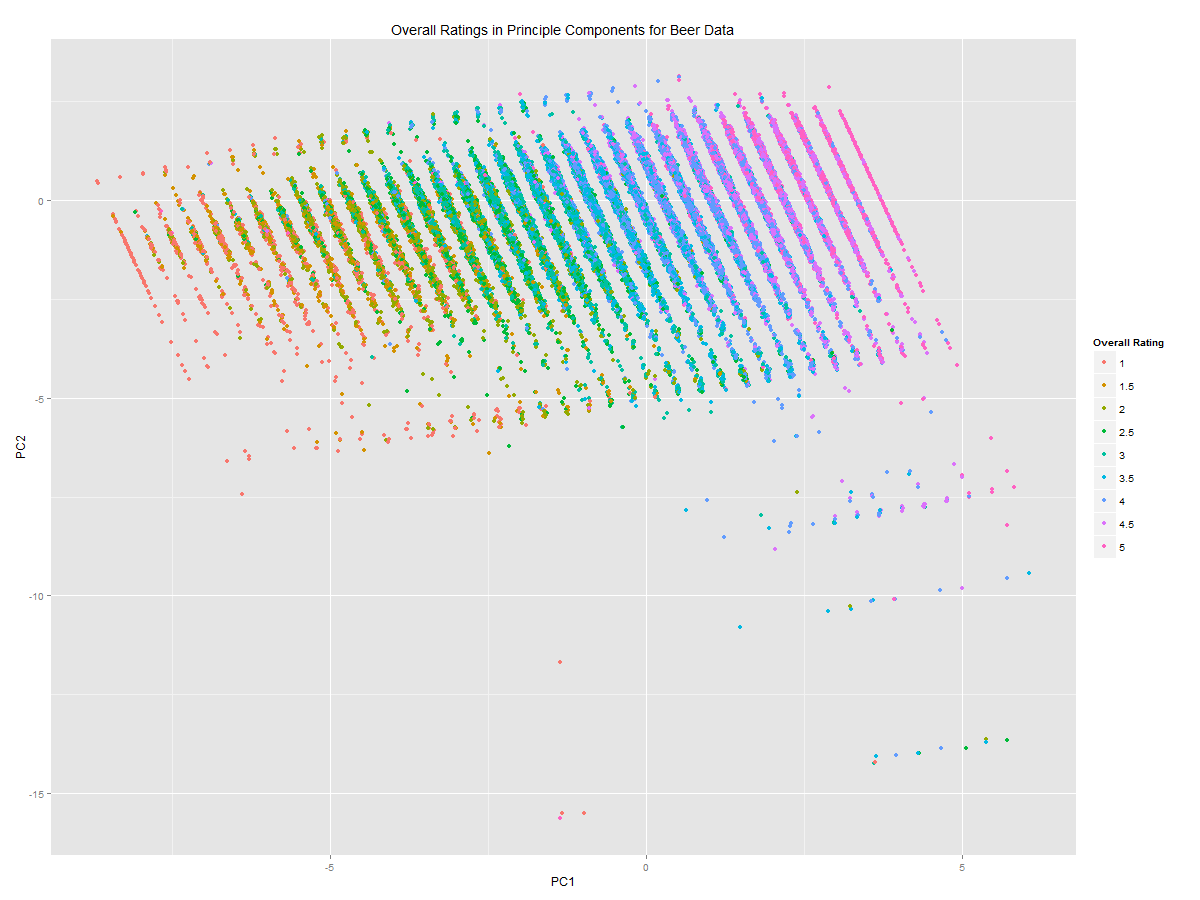
reviewAroma 0.4891178 0.06776329 -0.186259686 0.74453672 -0.40883284

reviewPalate 0.4881409 0.17404576 -0.226135782 -0.66271550 -0.49101663

reviewTaste 0.5034040 0.17210328 -0.370774753 -0.02128444 0.76094544

Mardia criterion for PCA1: 0.35238 [fulfilled by Appearance, Aroma, Palate and Taste]  
Mardia criterion for PCA2: 0.121832 [fulfilled by ABV, Appearance, Palate and Taste]

Kaiser criterion suggests using only PC1 (as sd is > 1) however PC2 is very close to the cut off value



The overall ratings are strongly dependent on PC1 but also increase slightly as score on PC2 increases.

PCA for Ales, Lagers and Stouts

20% sampled data, cleaned and standardised, pulled out all obs for ales, lagers and stouts  
Number of Ales=96024, Lagers=22084, Stouts=34977, Total=153085 (only accounts for half of the beers in our 20% sample)  
Variables = ABV, Appearance, Aroma, Palate, Taste, Rating

PCA Summary:  
Importance of components:

PC1 PC2 PC3 PC4 PC5 PC6

Standard deviation 1.9515 0.9379 0.71527 0.59839 0.51759 0.41767

Proportion of Variance 0.6347 0.1466 0.08527 0.05968 0.04465 0.02907

Cumulative Proportion 0.6347 0.7813 0.86660 0.92628 0.97093 1.00000

Rotations (Principle Components):  
 PC1 PC2 PC3 PC4 PC5 PC6

ABV 0.2412693 0.91665947 -0.20866892 0.15152750 -0.1650915 -0.08812527

reviewAppearance 0.3906596 0.03662312 0.88037215 0.15545005 -0.2137594 0.03362925

reviewAroma 0.4367883 0.05580435 0.02772638 -0.81455022 0.2922907 -0.23750193

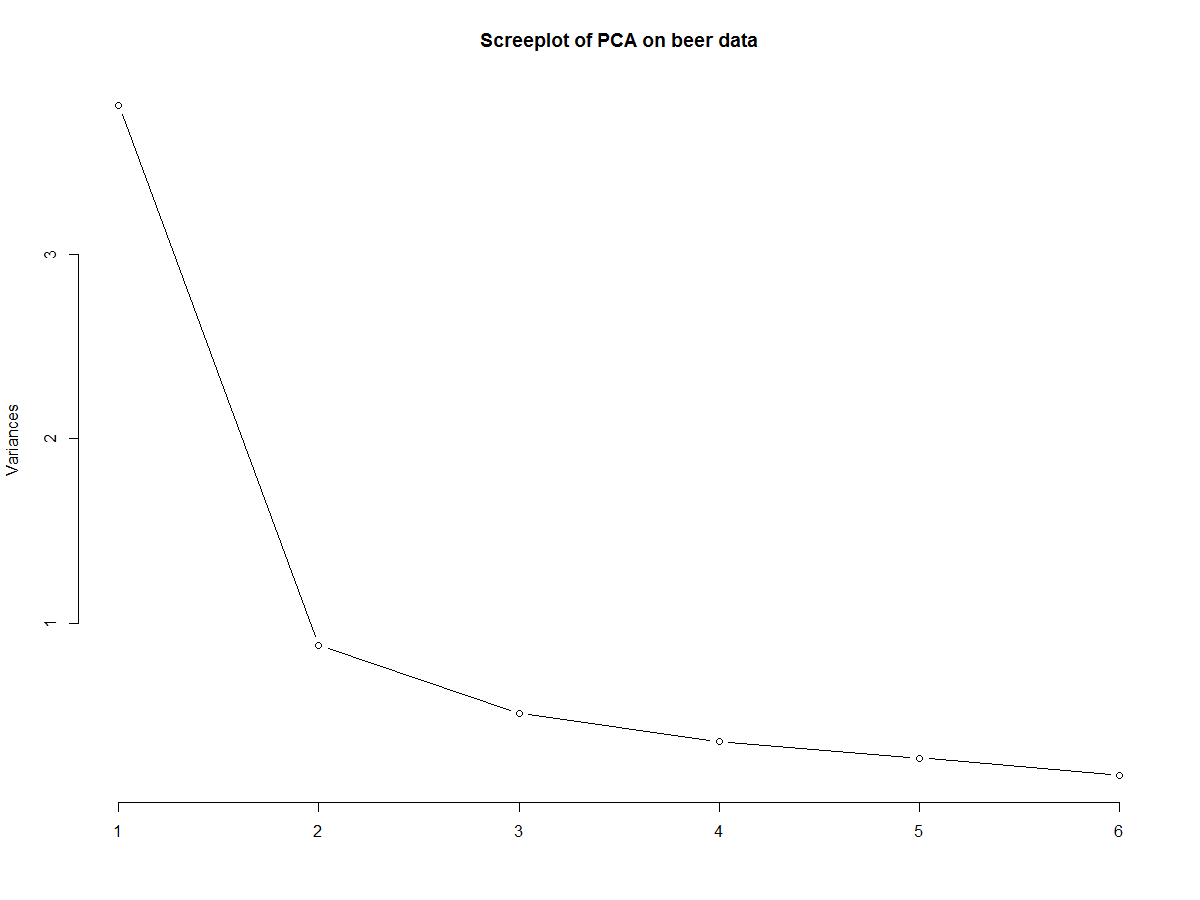
reviewPalate 0.4446159 -0.10506665 -0.10588290 0.51012897 0.7155707 -0.08828094

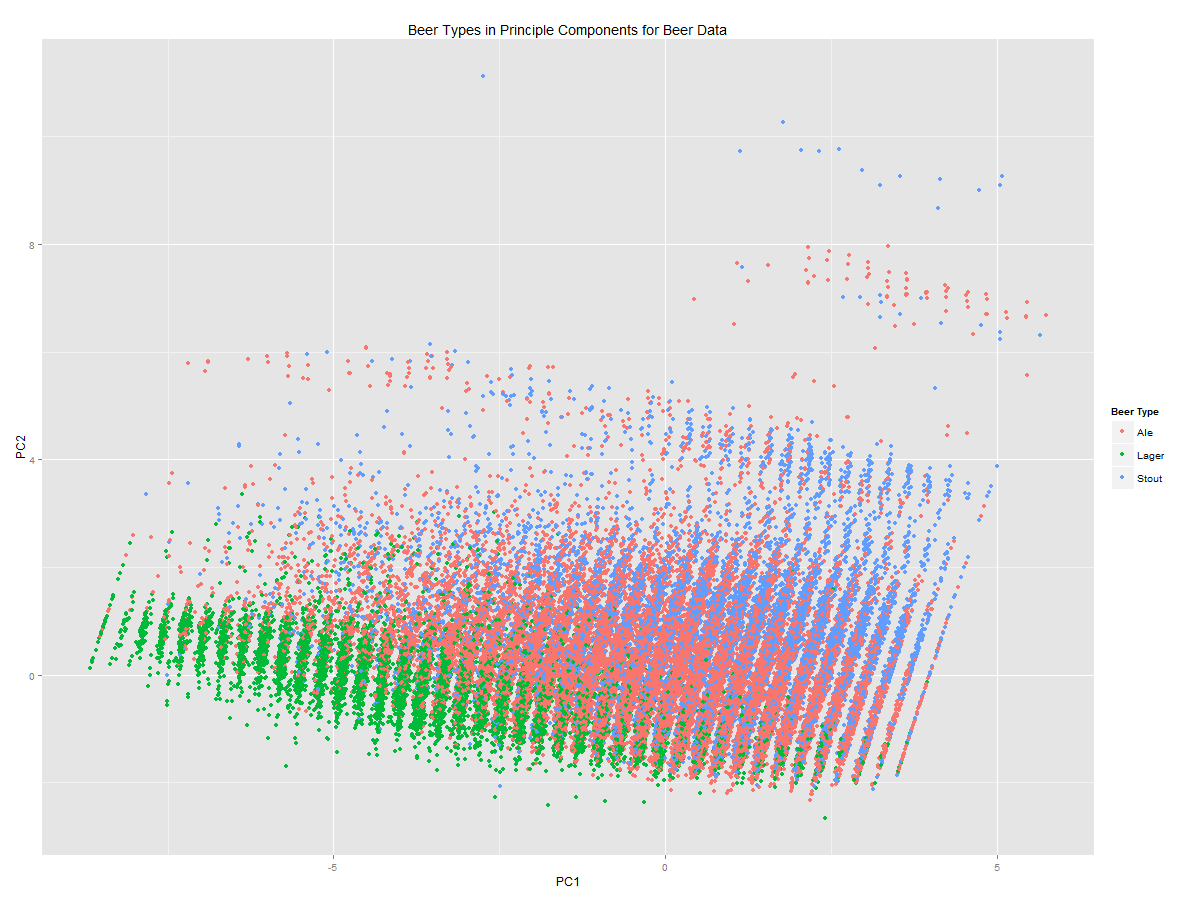
reviewTaste 0.4649560 -0.12805231 -0.26605125 -0.07400148 -0.1945949 0.80826474

rating 0.4295613 -0.35755230 -0.31406839 0.15386632 -0.5401011 -0.52307784

Mardia criterion for PCA1: 0.3254692 [fulfilled by Appearance, Aroma, Palate, Taste and Rating]  
Mardia criterion for PCA2: 0.6416616 [fulfilled by ABV]

Kaiser criterion suggests using only PC1 (as sd is > 1) however PC2 is very close to the cut off value





PCA Rotations for Ales only:  
 PC1 PC2 PC3 PC4 PC5

ABV 0.1921986 -0.93479268 0.14195134 -0.175192231 0.1540924

reviewAppearance 0.3629956 -0.03550017 -0.91682998 -0.042220447 0.1535909

reviewAroma 0.4336231 -0.08590640 0.07715690 0.816882824 -0.3270020

reviewPalate 0.4525818 0.08811622 0.08052092 -0.525265817 -0.7105485

reviewTaste 0.4815985 0.09484309 0.26870527 0.006865576 0.3313199

rating 0.4544199 0.31743155 0.23373760 -0.155808388 0.4807109

PCA Rotations for Lagers only:  
 PC1 PC2 PC3 PC4 PC5

ABV -0.1357680 0.970696541 0.1832490 0.03741535 -0.06288269

reviewAppearance -0.4095844 0.063735945 -0.7930232 0.38425617 -0.22469445

reviewAroma -0.4455286 0.003041432 -0.1937034 -0.81837275 0.05139826

reviewPalate -0.4536867 -0.057511020 0.1551898 0.32571214 0.79308013

reviewTaste -0.4680110 -0.094857041 0.2383227 -0.11556982 -0.08667958

rating -0.4364011 -0.203399431 0.4681148 0.24853280 -0.55355844

PCA Rotations for Stouts only:  
 PC1 PC2 PC3 PC4 PC5

ABV -0.1866903 -0.94182454 0.006617277 -0.17631182 0.1912972

reviewAppearance -0.3455552 0.07910015 -0.919092707 0.05269998 0.1571536

reviewAroma -0.4262977 -0.10918566 0.143736362 0.82885165 -0.2752881

reviewPalate -0.4599886 0.04115802 0.019219603 -0.49101215 -0.7353411

reviewTaste -0.4891908 0.08290437 0.275105230 -0.05351688 0.2829801

rating -0.4616400 0.29363426 0.241893334 -0.18757614 0.4920573

Mardia criterion Ale: PC1 retain Appearance, Aroma, Palate, Taste and Rating  
 PC2 retain ABV and Rating  
Mardia criterion Lager: PC1 retain ABV, Appearance, Aroma, Palate, Taste and Rating  
 PC2 retain ABV  
Mardia criterion Stout: PC1 retain ABV, Appearance, Aroma, Palate, Taste and Rating  
 PC2 retain ABV and Rating

PCA for American, Belgian and English beers

20% sampled data, cleaned and standardised, pulled out all obs for american, belgian and english  
Number of American=122442, Belgian=21026, English=19804, Total=163272 (only accounts for half of the beers in our 20% sample)  
Variables = ABV, Appearance, Aroma, Palate, Taste, Rating

PCA Summary:  
Importance of components:

PC1 PC2 PC3 PC4 PC5 PC6

Standard deviation 1.9016 0.9675 0.75356 0.63112 0.53926 0.43688

Proportion of Variance 0.6027 0.1560 0.09464 0.06639 0.04847 0.03181

Cumulative Proportion 0.6027 0.7587 0.85334 0.91972 0.96819 1.00000

PCA Rotations (Principle Components):  
 PC1 PC2 PC3 PC4 PC5 PC6

ABV 0.2066293 -0.92746401 -0.199839900 0.13730688 -0.1693858 -0.09815366

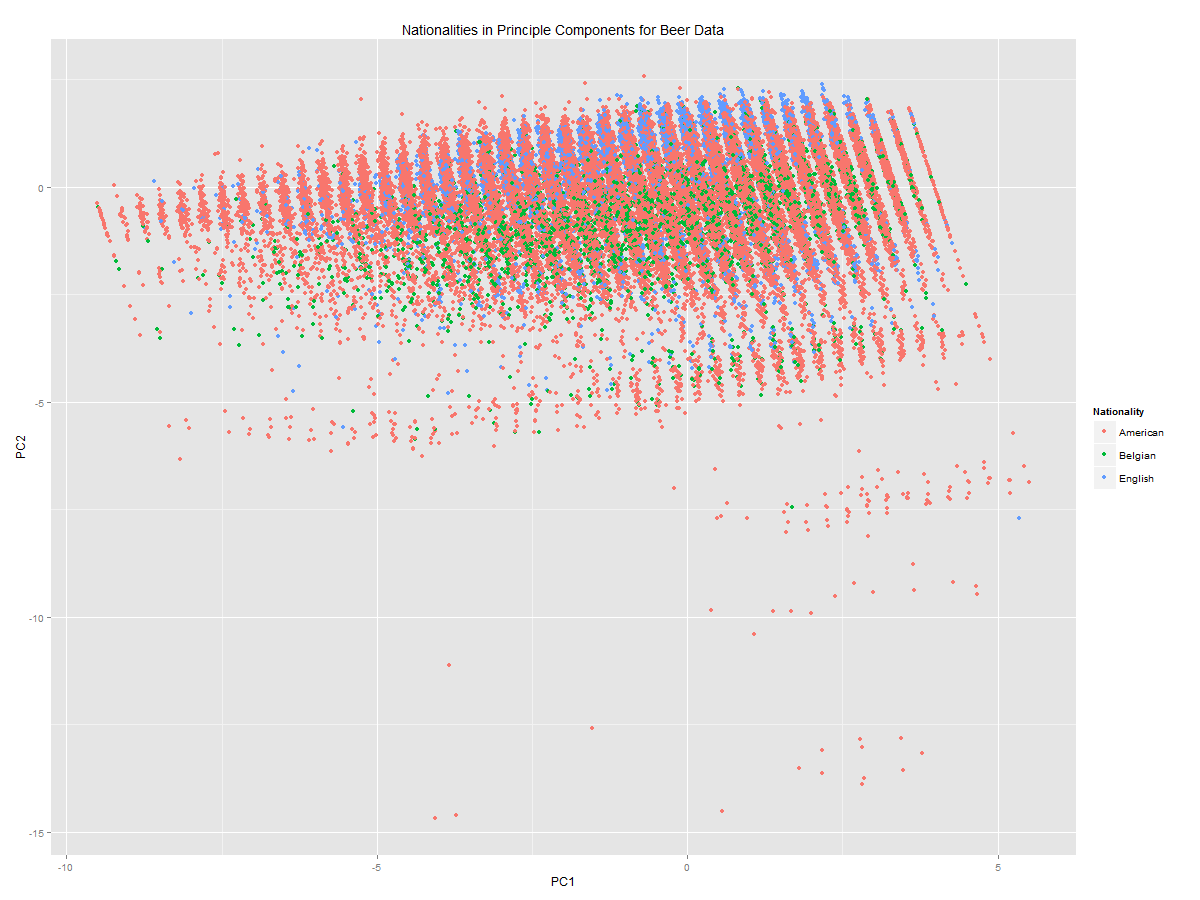
reviewAppearance 0.3816419 -0.05742328 0.893637177 0.12104990 -0.1924268 0.02799074

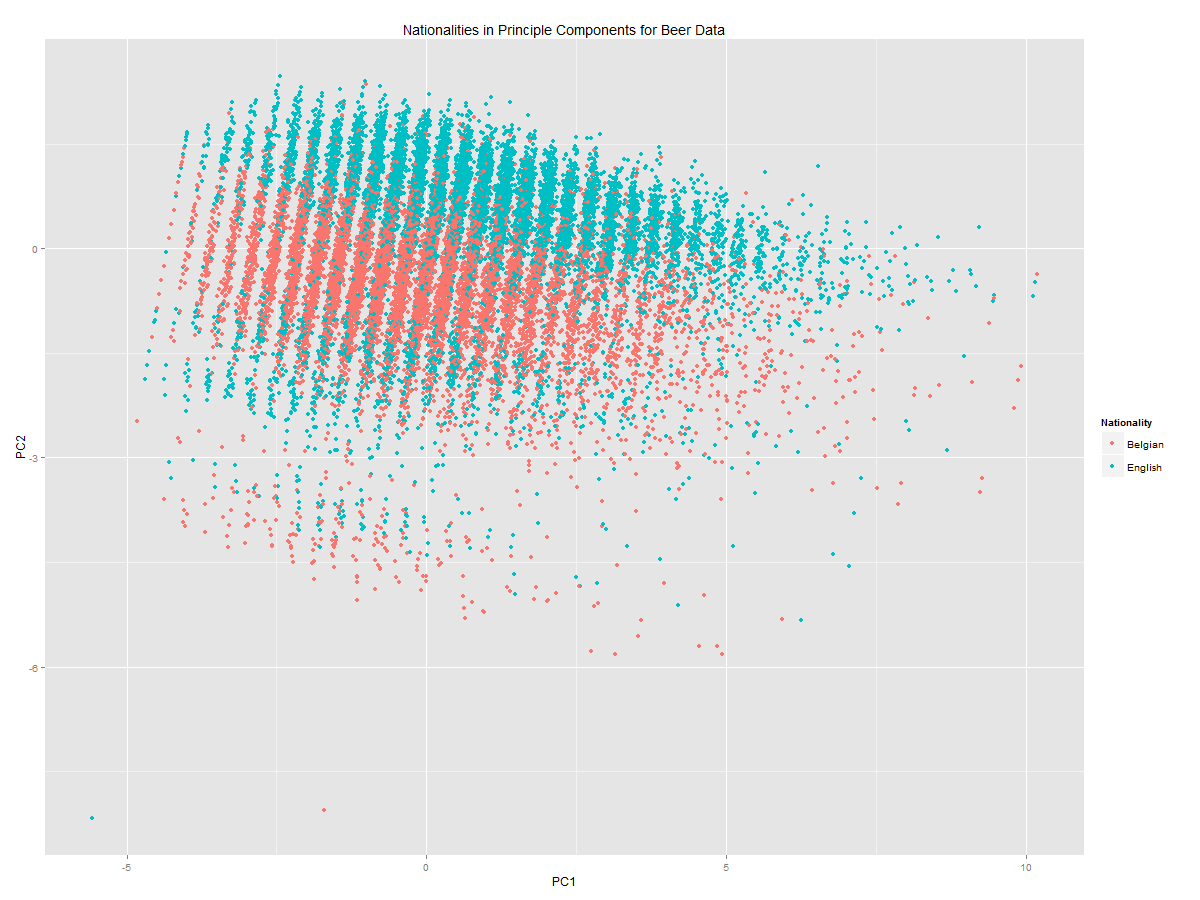
reviewAroma 0.4390472 -0.05888104 -0.004674512 -0.80831641 0.3280675 -0.20674865

reviewPalate 0.4493354 0.07929436 -0.106892079 0.53539644 0.7011115 -0.04666371

reviewTaste 0.4723855 0.11845675 -0.267063602 -0.05765219 -0.2604136 0.78762807

rating 0.4402738 0.33574766 -0.280545675 0.15213547 -0.5169908 -0.56947511





Further work on beer nationalities:

Used full data set and defined beers that are American (n=641174), Belgian (n=106962), English (n=102485), German (n=21961), Scottish (n=11763), Irish (n=20086), and Russian (n=53555).

Done PCA analysis on all beers with defined nationality using ABV, Appearance, Aroma, Palate, Taste, Rating.

Importance of components:

PC1 PC2 PC3 PC4 PC5 PC6

Standard deviation 1.9034 0.9632 0.75277 0.6326 0.53962 0.43754

Proportion of Variance 0.6038 0.1546 0.09444 0.0667 0.04853 0.03191

Cumulative Proportion 0.6038 0.7584 0.85286 0.9196 0.96809 1.00000

PC1 PC2 PC3 PC4 PC5 PC6

ABV -0.2159464 -0.92144526 -0.20500078 0.14699021 -0.1763113 -0.09792173

reviewAppearance -0.3810114 -0.06225998 0.89728253 0.09612842 -0.1886699 0.03163925

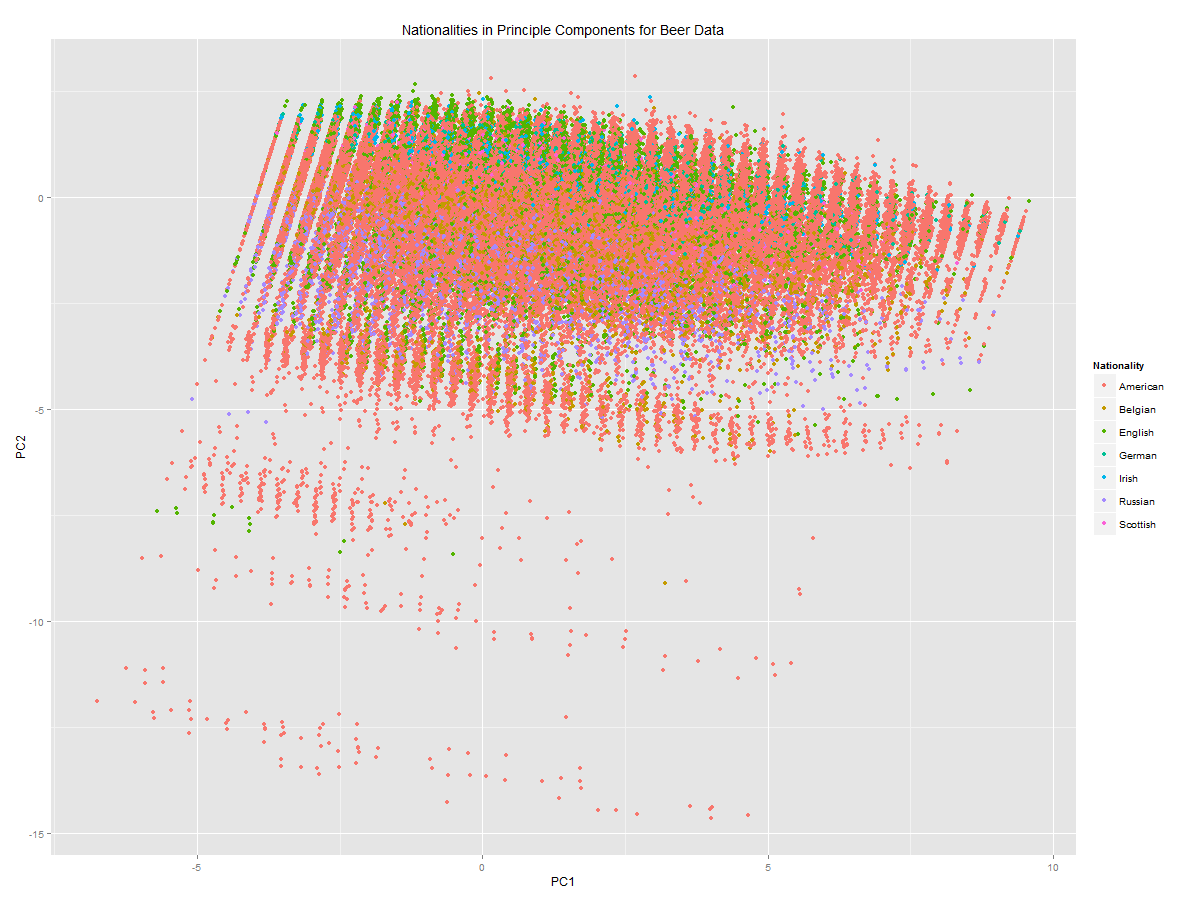
reviewAroma -0.4381681 -0.05889675 -0.02826221 -0.80890679 0.3218473 -0.21406829

reviewPalate -0.4488257 0.08338894 -0.09247360 0.53560340 0.7017659 -0.06086137

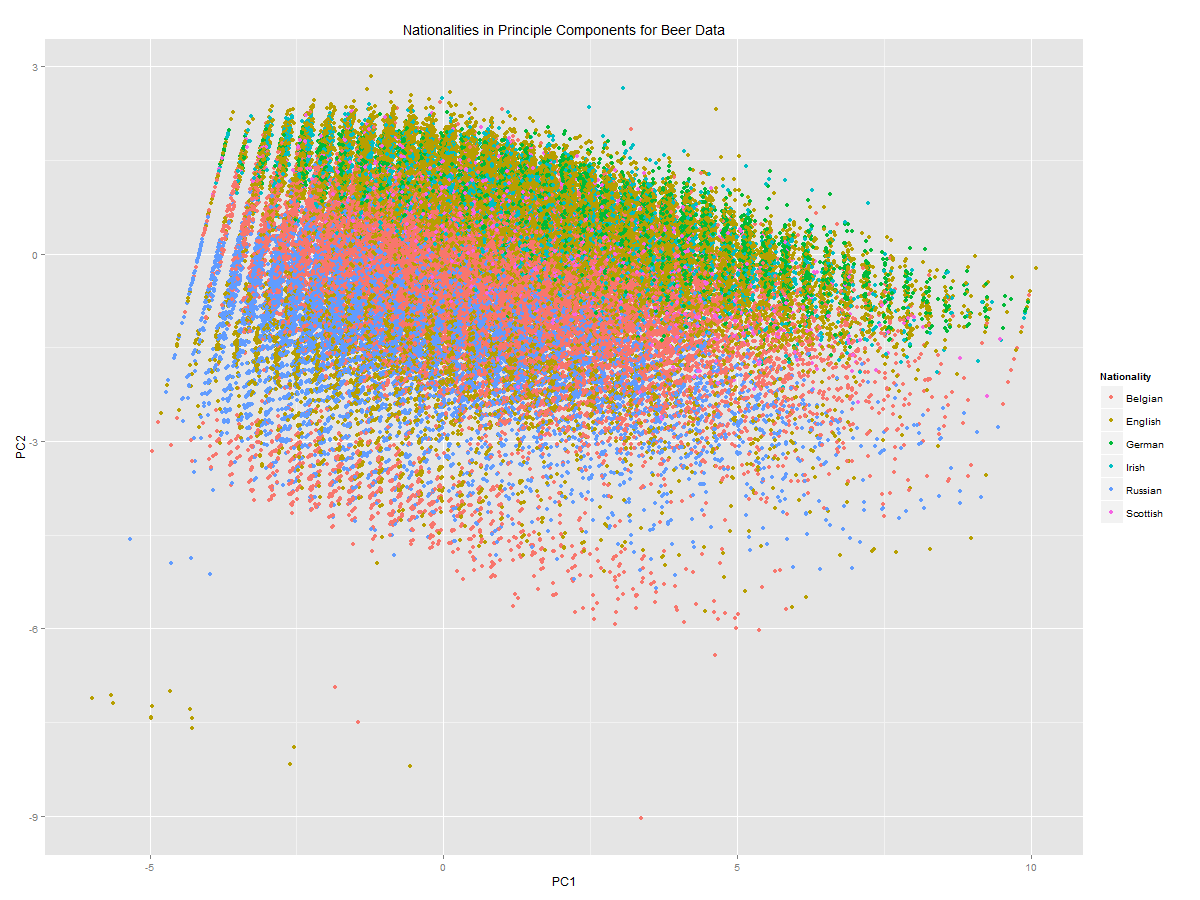
reviewTaste -0.4719599 0.12390022 -0.26558193 -0.05103756 -0.2438164 0.79329537

rating -0.4381957 0.34826503 -0.27013860 0.15920798 -0.5270774 -0.55728216

Plot of data on PC1 and PC2, coloured by nationality. It is difficult to see any patterns as there are so many American beers:



Without American beers you can see a pattern strongly dependent on alcohol levels:



Conducted an anova on pc1 scores for the beer nationalities, and used tukey multiple comparisons:

Df Sum Sq Mean Sq F value Pr(>F)

beer.data.n$Nationality 6 142139 23690 6831 <2e-16 \*\*\*

Residuals 917621 3182212 3

diff lwr upr p adj

Belgian-American -0.3556591 -0.37395541 -0.33736277 0.0000000

English-American 0.4906232 0.47176247 0.50948395 0.0000000

German-American 1.0903358 1.05158238 1.12908929 0.0000000

Irish-American 0.6818487 0.64110590 0.72259153 0.0000000

Russian-American -1.1412407 -1.16613381 -1.11634753 0.0000000

Scottish-American 0.5137181 0.46123525 0.56620098 0.0000000

English-Belgian 0.8462823 0.82195820 0.87060641 0.0000000

German-Belgian 1.4459949 1.40430841 1.48768143 0.0000000

Irish-Belgian 1.0375078 0.99396574 1.08104987 0.0000000

Russian-Belgian -0.7855816 -0.81483227 -0.75633089 0.0000000

Scottish-Belgian 0.8693772 0.81469278 0.92406163 0.0000000

German-English 0.5997126 0.55777532 0.64164992 0.0000000

Irish-English 0.1912255 0.14744327 0.23500774 0.0000000

Russian-English -1.6318639 -1.66147089 -1.60225687 0.0000000

Scottish-English 0.0230949 -0.03178094 0.07797075 0.8782098

Irish-German -0.4084871 -0.46383137 -0.35314286 0.0000000

Russian-German -2.2315765 -2.27655069 -2.18660231 0.0000000

Scottish-German -0.5766177 -0.64109610 -0.51213934 0.0000000

Russian-Irish -1.8230894 -1.86978869 -1.77639009 0.0000000

Scottish-Irish -0.1681306 -0.23382389 -0.10243732 0.0000000

Scottish-Russian 1.6549588 1.59772855 1.71218901 0.0000000

The only non-significant relationship is between Scottish and English beers (i.e. they show no significant differences in aroma, palate, taste and rating). The largest difference is between Russian and German beers, where German beers score highly on average on pc1 and Russian beers score low on average.

An anova conducted on pc2:

Df Sum Sq Mean Sq F value Pr(>F)

beer.data.n$Nationality 6 94499 15750 19096 <2e-16 \*\*\*

Residuals 917621 756824 1

diff lwr upr p adj

Belgian-American -0.33993415 -0.34885686 -0.331011442 0.0000000

English-American 0.39194300 0.38274504 0.401140963 0.0000000

German-American 0.78039926 0.76150006 0.799298459 0.0000000

Irish-American 0.76079181 0.74092244 0.780661175 0.0000000

Russian-American -0.87668145 -0.88882128 -0.864541616 0.0000000

Scottish-American 0.52996221 0.50436748 0.555556935 0.0000000

English-Belgian 0.73187715 0.72001482 0.743739475 0.0000000

German-Belgian 1.12033341 1.10000382 1.140662993 0.0000000

Irish-Belgian 1.10072596 1.07949146 1.121960457 0.0000000

Russian-Belgian -0.53674730 -0.55101221 -0.522482385 0.0000000

Scottish-Belgian 0.86989636 0.84322798 0.896564734 0.0000000

German-English 0.38845626 0.36800437 0.408908149 0.0000000

Irish-English 0.36884881 0.34749719 0.390200431 0.0000000

Russian-English -1.26862445 -1.28306313 -1.254185764 0.0000000

Scottish-English 0.13801921 0.11125748 0.164780937 0.0000000

Irish-German -0.01960745 -0.04659762 0.007382714 0.3279331

Russian-German -1.65708071 -1.67901362 -1.635147791 0.0000000

Scottish-German -0.25043705 -0.28188173 -0.218992381 0.0000000

Russian-Irish -1.63747325 -1.66024747 -1.614699043 0.0000000

Scottish-Irish -0.23082960 -0.26286676 -0.198792446 0.0000000

Scottish-Russian 1.40664365 1.37873374 1.434553564 0.0000000

The only non-significant relationship is between Irish and German beers (i.e. they have similar alcohol levels). Once again the largest difference is between German and Russian beers, where German beers contain low alcohol levels and Russian beers contain high alcohol levels (as expected given that Russian beers score lowest on PC1 etc).

Further work on beer types:

Run on full data set, found all ales (n= 502388), lagers (n= 118289) and stouts (n= 180253)  
PCA results:

Importance of components:

PC1 PC2 PC3 PC4 PC5 PC6

Standard deviation 1.953 0.9373 0.7133 0.59696 0.51639 0.41679

Proportion of Variance 0.636 0.1464 0.0848 0.05939 0.04444 0.02895

Cumulative Proportion 0.636 0.7824 0.8672 0.92661 0.97105 1.00000

PC1 PC2 PC3 PC4 PC5 PC6

ABV -0.2414827 -0.91700301 0.20890203 -0.14835943 0.1670279 0.08510389

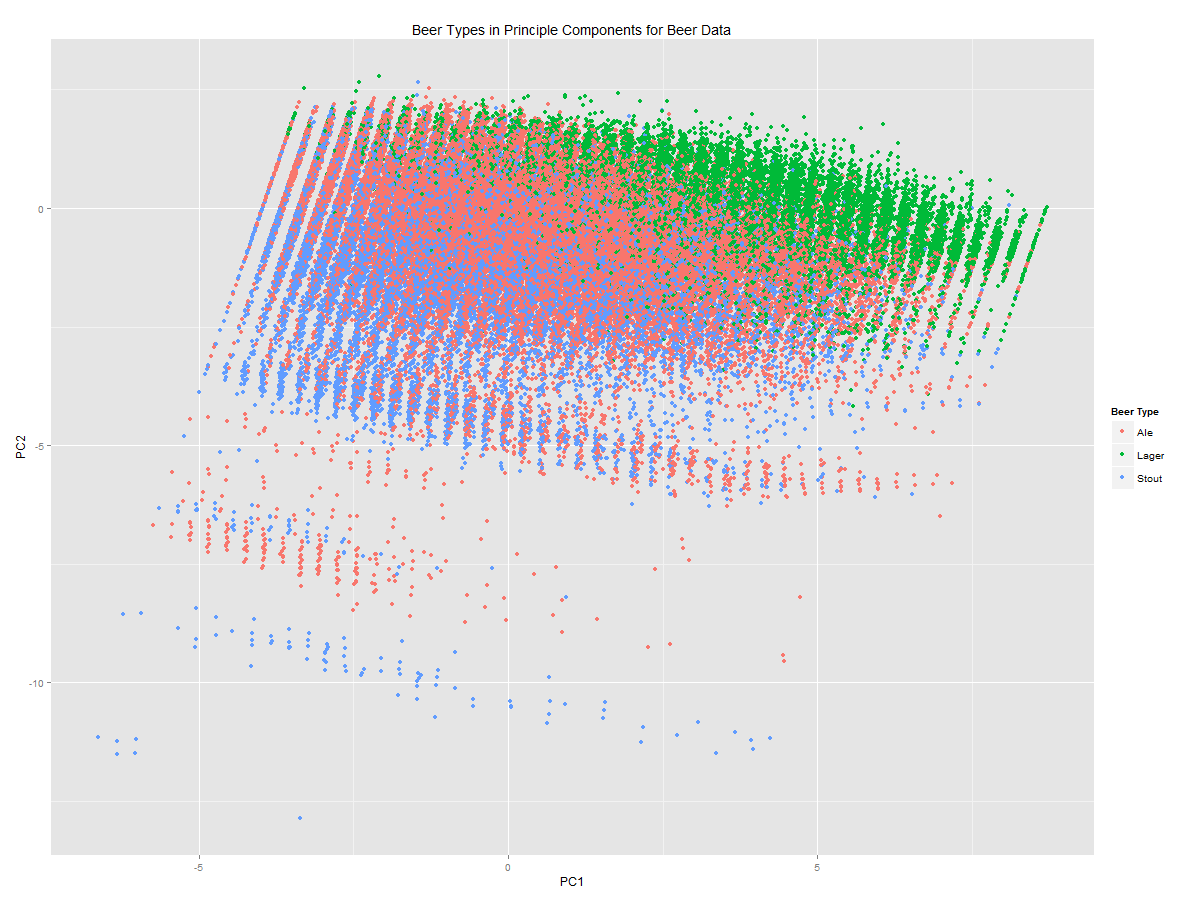
reviewAppearance -0.3910529 -0.03757230 -0.88101236 -0.14335059 0.2189562 -0.03149535

reviewAroma -0.4371121 -0.05168844 -0.01828632 0.81120218 -0.2968711 0.24442889

reviewPalate -0.4444738 0.10315689 0.09768550 -0.51948052 -0.7092306 0.09690786

reviewTaste -0.4650152 0.12777881 0.26279034 0.07443743 0.1806611 -0.81252405

rating -0.4288364 0.35784542 0.31818478 -0.15488759 0.5480701 0.51228256



PC1 Anova results (w/ tukey comparisons):

Df Sum Sq Mean Sq F value Pr(>F)

beer.data.type$Type 2 704244 352122 121660 <2e-16 \*\*\*

Residuals 764121 2211607 3

diff lwr upr p adj

Lager-Ale 2.3484267 2.3351616 2.3616919 0

Stout-Ale -0.7935706 -0.8047248 -0.7824164 0

Stout-Lager -3.1419973 -3.1572915 -3.1267031 0

All have significant differences on PC1. Lager and stout differ most with lager being the most liked (highest on PC1) and stout being least favoured. Ale is closer to stout than lager on PC1.

Pc2 anova results:

Df Sum Sq Mean Sq F value Pr(>F)

beer.data.type$Type 2 52181 26090 32202 <2e-16 \*\*\*

Residuals 764121 619106 1

diff lwr upr p adj

Lager-Ale 0.3198147 0.3127963 0.3268332 0

Stout-Ale -0.5044709 -0.5103724 -0.4985693 0

Stout-Lager -0.8242856 -0.8323776 -0.8161936 0

Again all are significantly different from each other. The biggest difference is, as expected, between stout and lager, where stout has the lowest score on PC2 (and therefore the most alcohol content) and lager has the highest (lowest alcohol). Lager and ale are more similar in terms of alcohol content, but not by too much.

PCA on full data set

Variables: ABV, Appearance, Aroma, Palate, Taste, and Rating

Importance of components:

PC1 PC2 PC3 PC4 PC5 PC6

Standard deviation 1.9161 0.9581 0.74097 0.62615 0.53263 0.43097

Proportion of Variance 0.6119 0.1530 0.09151 0.06534 0.04728 0.03096

Cumulative Proportion 0.6119 0.7649 0.85642 0.92176 0.96904 1.00000

PC1 PC2 PC3 PC4 PC5 PC6

ABV 0.2141756 0.93123660 0.189148926 -0.13954822 0.1550273 -0.08742257

reviewAppearance 0.3847739 0.04591445 -0.893441923 -0.12465665 0.1874952 0.03014459

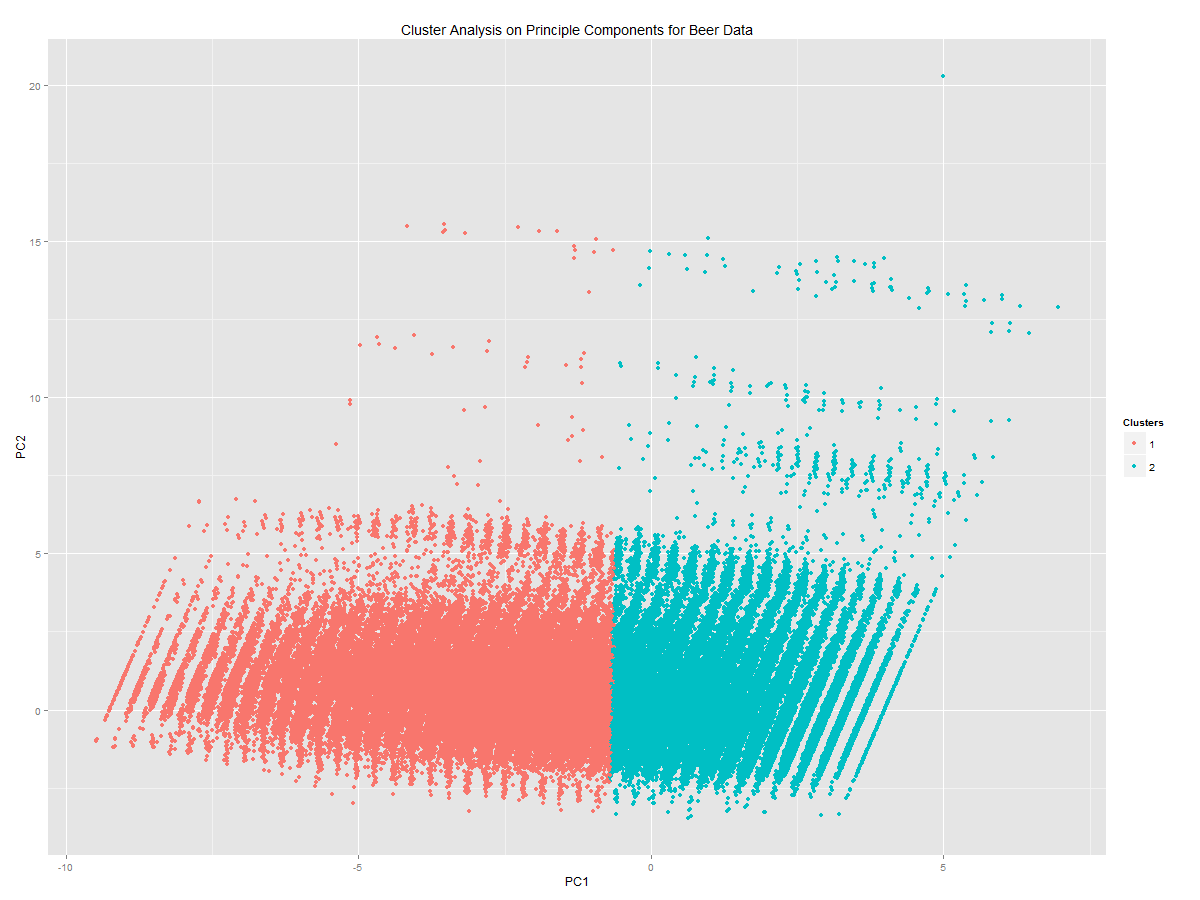
reviewAroma 0.4367273 0.04914228 0.007692093 0.82189892 -0.2903324 -0.21675873

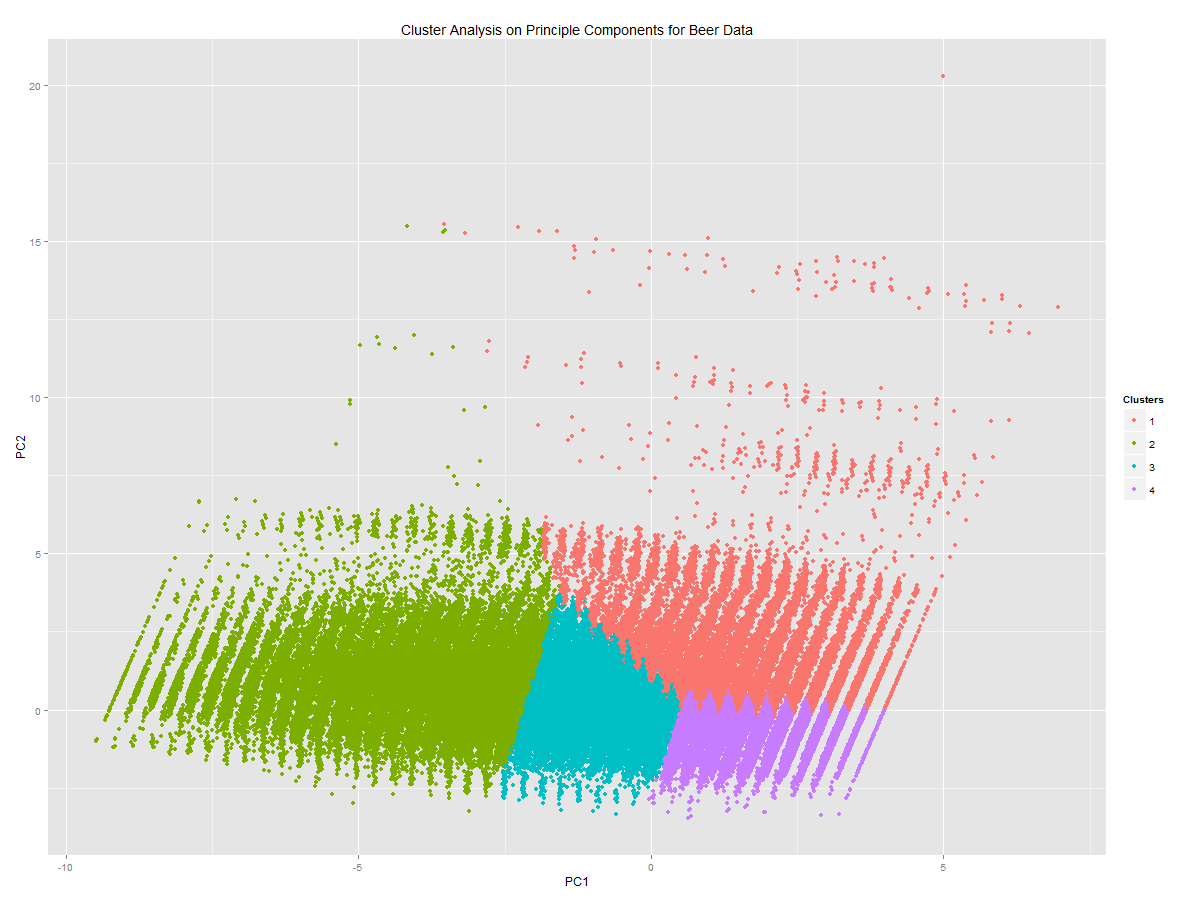
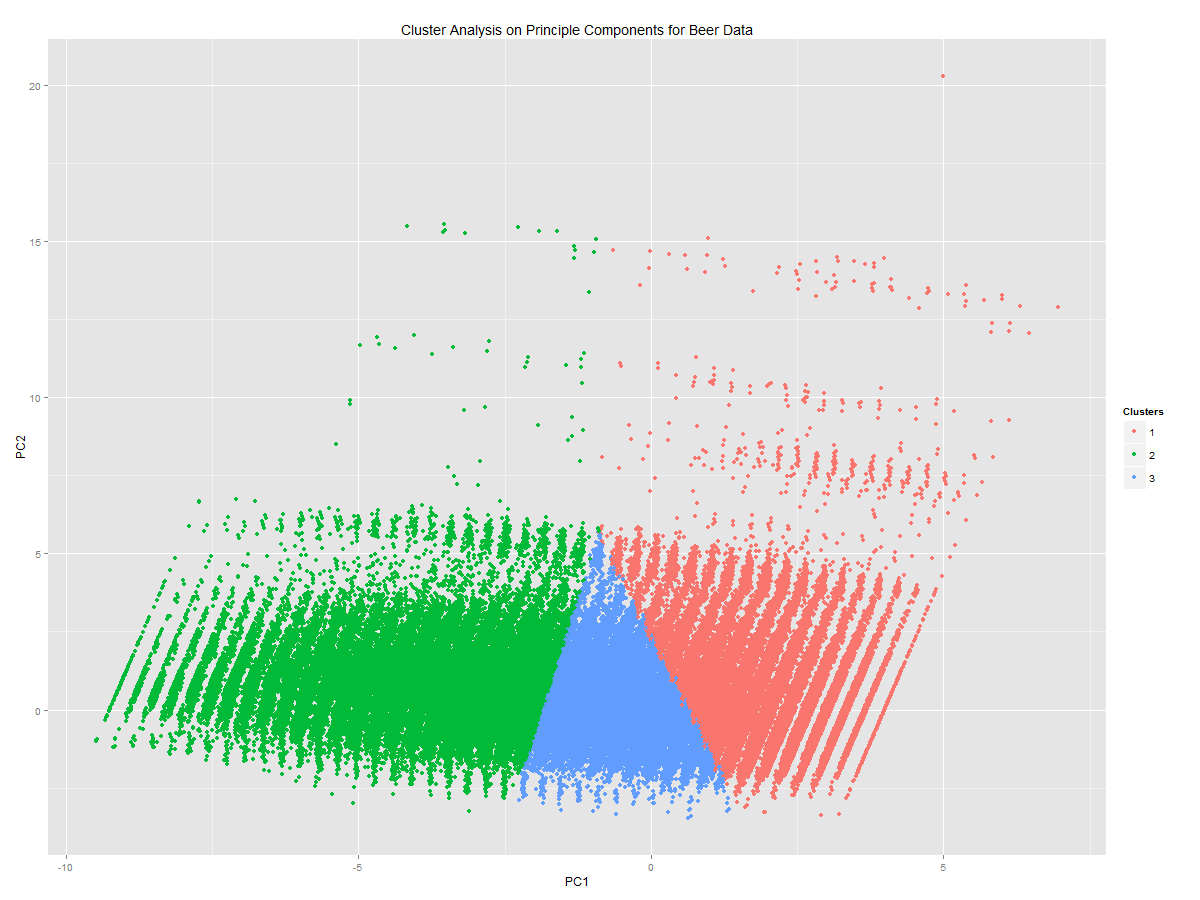
reviewPalate 0.4483970 -0.08722945 0.105871828 -0.50674744 -0.7200605 -0.06958569

reviewTaste 0.4706472 -0.11953695 0.266221117 0.04809865 0.2358234 0.79712128

rating 0.4390654 -0.32615524 0.289555726 -0.17424903 0.5314314 -0.55156119

Cluster analysis on PCA





Means table for clusters on PC1 and PC2:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Clusters | PC1 1 | PC2 1 | PC1 2 | PC2 2 | PC1 3 | PC2 3 | PC1 4 | PC2 4 |
| 2 | 0.9753569 | -0.008268 | -2.313956 | 0.0196166 |  |  |  |  |
| 3 | -3.419725 | 0.2403275 | -0.329303 | -0.298902 | 1.624691 | 0.238189 |  |  |
| 4 | 1.518209 | 1.101513 | -0.663052 | -0.239378 | -3.731278 | 0.2750871 | 1.333596 | -0.610603 |