Tim Miller
15.071 | HW5

Problem 1a

To understand the importance of variables we summed across all columns. The columns with the highest values were the ones that were most important to the most customers.

```
# sum data set to get importance of columns
colSums(data)
```

```
driving_properties          interior        technology           comfort       reliability          handling             power       consumption            sporty
               547               195               348               308               325               231               388               203               323
            safety            gender         household
               317                49               469
```

Excluding the gender and household columns, we find the ranking of the features from most important to least important (based on total number of "important" votes received)

- driving_properties
- power
- technology
- reliability
- sporty
- safety
- comfort
- handling
- consumption
- interior

Then looking at the gender and household columns, we find that the survey consisted of ~94% males and ~6% females.

```
#calculate gender breakdown
percent_female = sum(data$gender) / nrow(data)
percent_male = 1 - percent_female

percent_male #0.9382093
percent_female #0.06179067
```

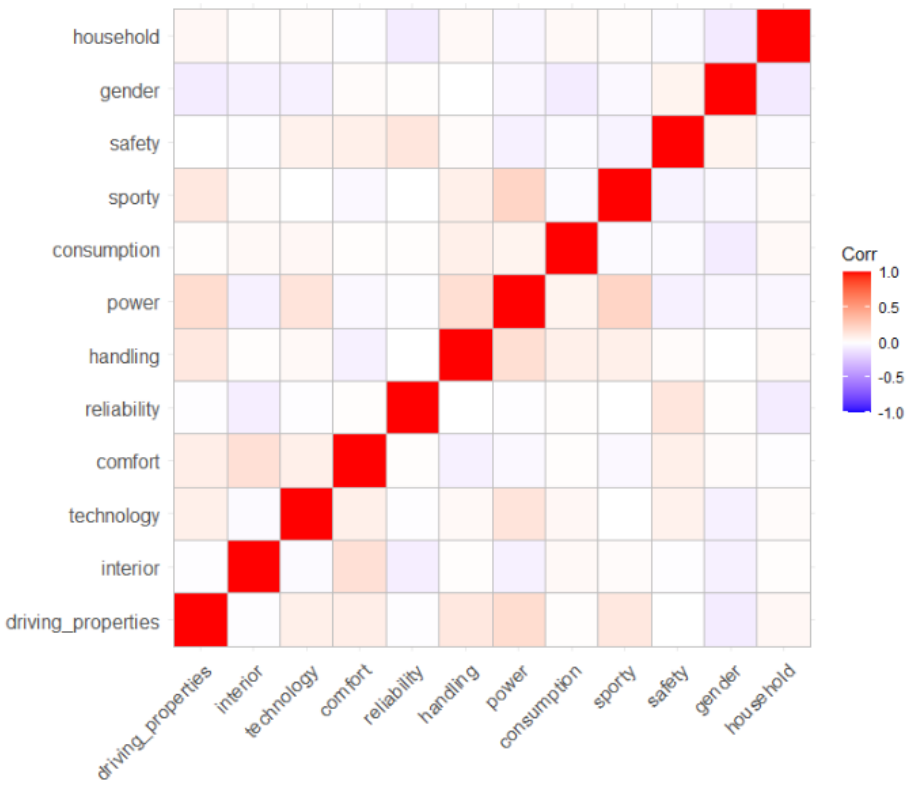
We find that ~60% of respondents have at least 3 people and the remaining have 1-2 people.

```
#calculate household breakdown
percent_3people = sum(data$household) / nrow(data)
percent_1_2people = 1 - percent_3people

percent_3people #0.591425
percent_1_2people #0.408575
```

Problem 1b

```
correlation = cor(data)
ggcorrplot(correlation)
```



There are a few interesting relationships in the data:

- driving_properties is has a positive relationship with technology, comfort, handling, power, and sporty. This makes sense because the latter three are typical inputs into overall driving performance
- safety and reliability are correlated, which is intuitive. One might expect a reliable car also to be safe.
- sporty and power have a strong correlation. Again this is intuitive: a sporty car probably is powerful.
- Interestingly there is a negative correlation across gender and most of the variables. Something to continue to explore.

Problem 1c

Normalizing data is most helpful when there are scaling issues across the data we are comparing. Oftne the issue is that distance between points we are analyzing is driven by variables with the larger scale, which skews the data. But since the data is all binary here, we do not have a scaling problem, therefore fine if we do not normalize.

Problem 1d

We use the following to calculate cluster size, finding 363 respondents in group 1 and 430 respondents in group 2.

```
> table(assignments)
assignments
  1   2
363 430
```

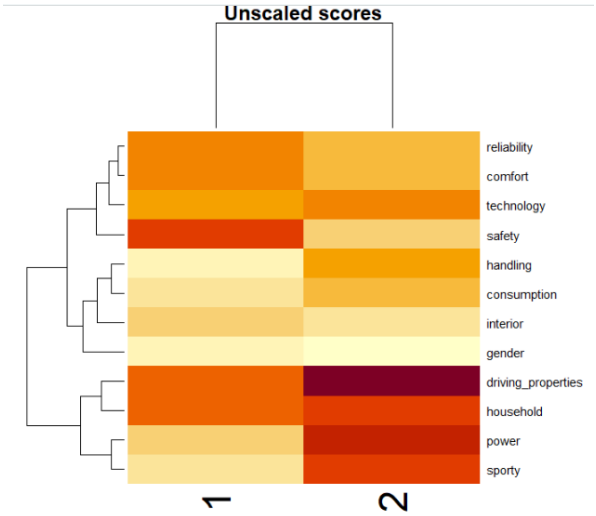
Then, printing out the summary of the clustersMeans

```
> clusterMeans
                            1          2
driving_properties 0.50964187 0.84186047
interior           0.29201102 0.20697674
technology         0.41597796 0.45813953
comfort            0.43801653 0.34651163
reliability        0.49035813 0.34186047
handling           0.14600551 0.41395349
power              0.25344353 0.68837209
consumption        0.19283747 0.30930233
sporty             0.19283747 0.58837209
safety             0.57575758 0.25116279
gender             0.09917355 0.03023256
household          0.53994490 0.63488372
>
```

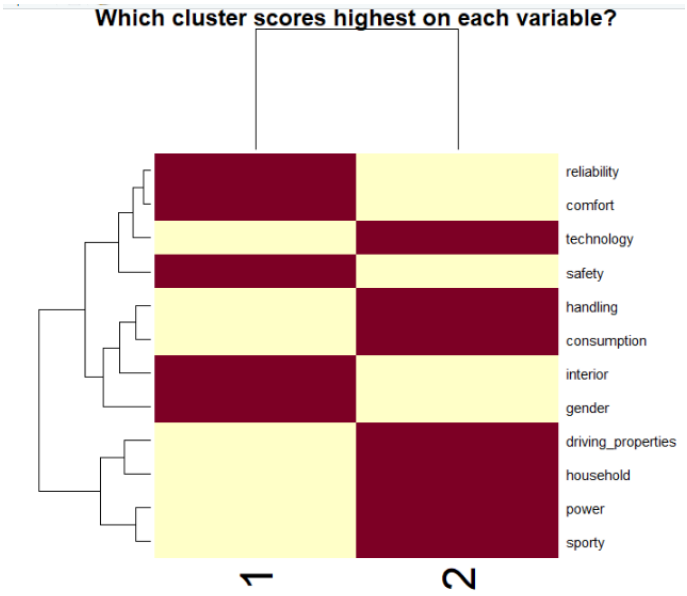
We find a few interesting takeaways:

- Cluster 1 has a slightly percentage of female respondents (higher mean value)
- Cluster 2 has slightly higher percentage of households with at least 3 people
- Cluster 2 puts more importance on driving_properties, power, sporty, consumption
- Cluster 1 puts more importance on reliability and safety
- Relatively equal importance for interior, technology, comfort

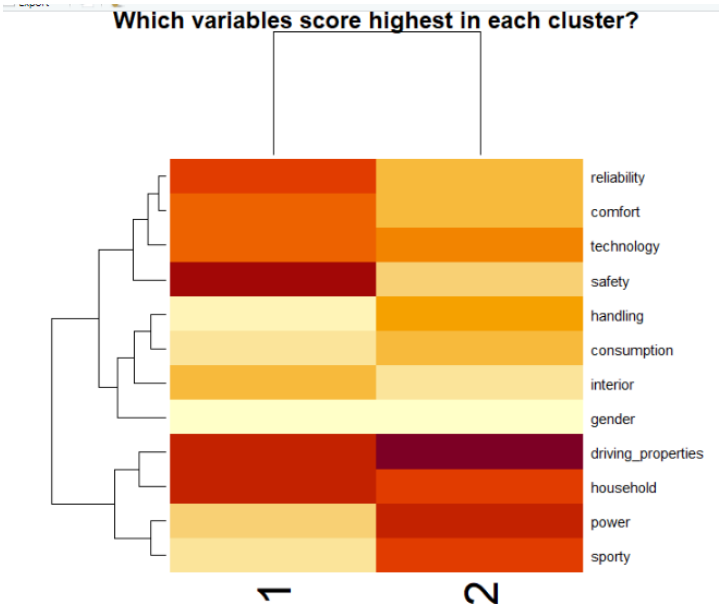
```
> heatmap(clusterMeans, scale="none", main="Unscaled scores")
```



```
> heatmap(clusterMeans, scale="row",
+         main="Which cluster scores highest on each variable?")
```




```
> heatmap(clusterMeans, scale="col",
+         main="Which variables score highest in each cluster?")
```



Problem 1e

We use the following to calculate cluster size, finding 207 respondents in group 1 and 430 respondents in group 2, and 156 respondents in group 3.

```
> table(assignments)
assignments
  1   2   3
207 430 156
```

From this, group 2 in problem 1d is clearly the summation of groups 1 and 3 in problem 1e. Cluster 2 in this problem is the same as the one in problem 1d.

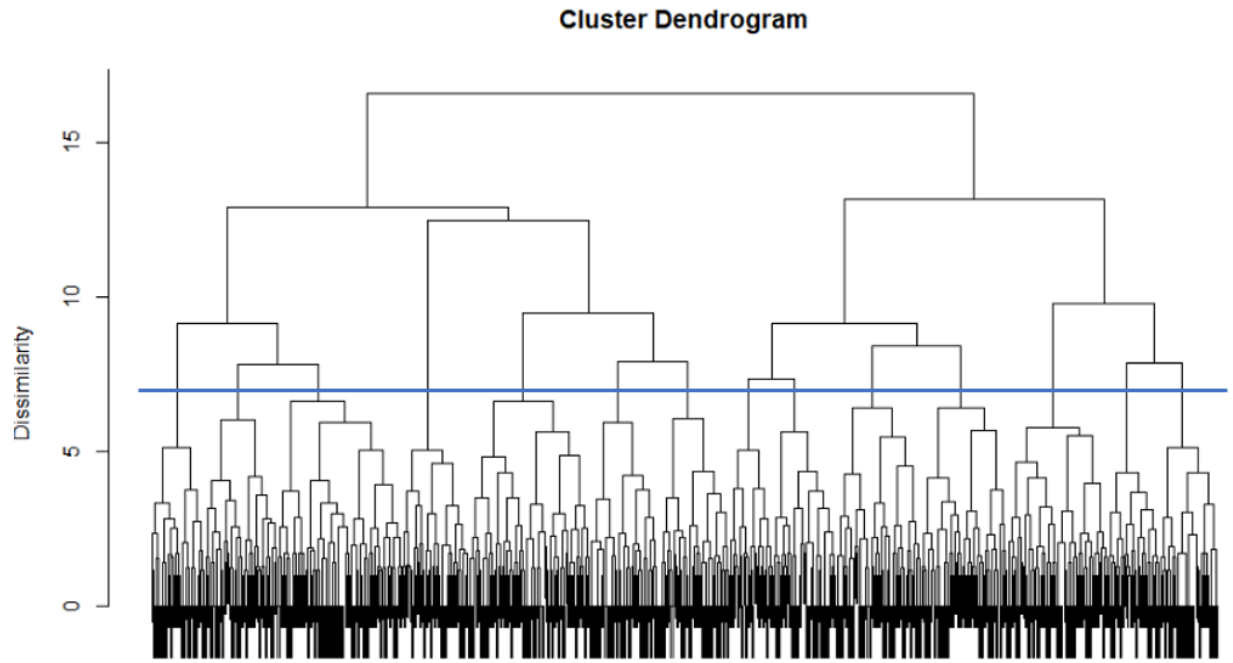
We use the following to analyze the clusters

```
> clusterMeans
                           1         2         3
driving_properties 0.6280193 0.84186047 0.35256410
interior           0.1787440 0.20697674 0.44230769
technology         0.4589372 0.45813953 0.35897436
comfort            0.4202899 0.34651163 0.46153846
reliability        0.6425121 0.34186047 0.28846154
handling           0.1594203 0.41395349 0.12820513
power              0.3816425 0.68837209 0.08333333
consumption        0.1352657 0.30930233 0.26923077
sporty             0.2850242 0.58837209 0.07051282
safety             0.8743961 0.25116279 0.17948718
gender             0.1159420 0.03023256 0.07692308
household          0.4396135 0.63488372 0.67307692
```

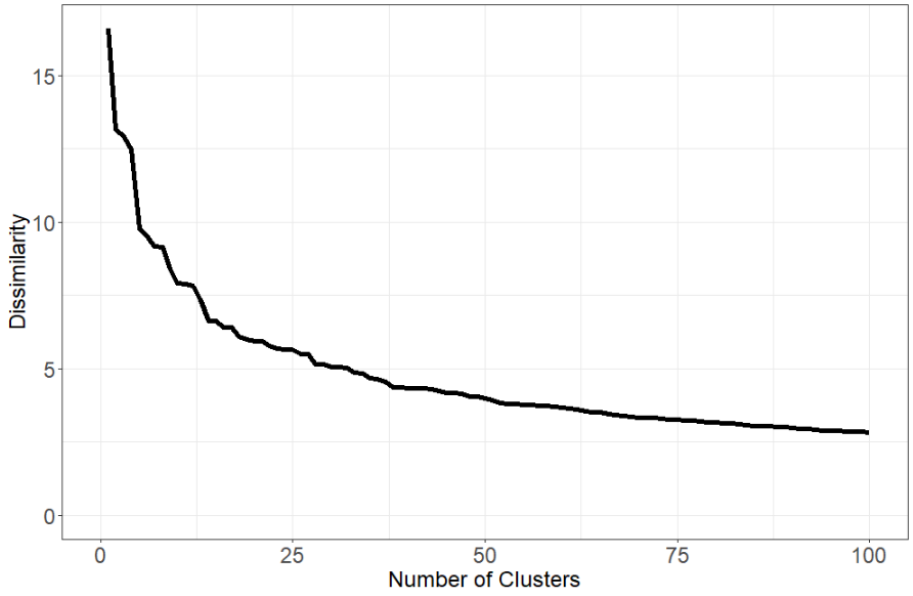
- Of the three clusters, cluster 1 has the higher portion of females, cluster 3 has second highest, and cluster 2 still has the lowest
- Cluster 1 cares more about driving_properties, sporty, and power relative to cluster 3, but both clusters care less about those properties relative to cluster 2
- Relative to cluster 1, cluster 3 cares much more about interior and slightly more about comfort
- Cluster 1 cares much more about safety compared to the other two clusters

Problem 1f

Dendogram



Scree plot



Looking at both the scree plot and the dendogram, my suggestion is we use 14 clusters. At 14 clusters, there is a reasonable amount of dissimilarity, but it is not significantly higher than the dissimilarity at 100 clusters. Furthermore, 14 clusters represents a reasonable number of clusters to work on without worrying too much about overfitting the clusters.

The scree plot above shows the blue line drawn across at 14 clusters.

Problem 1g

Number of respondents in each cluster below. Looks like clusters 3, 5, and 9 have larger number of respondents compared to other clusters.

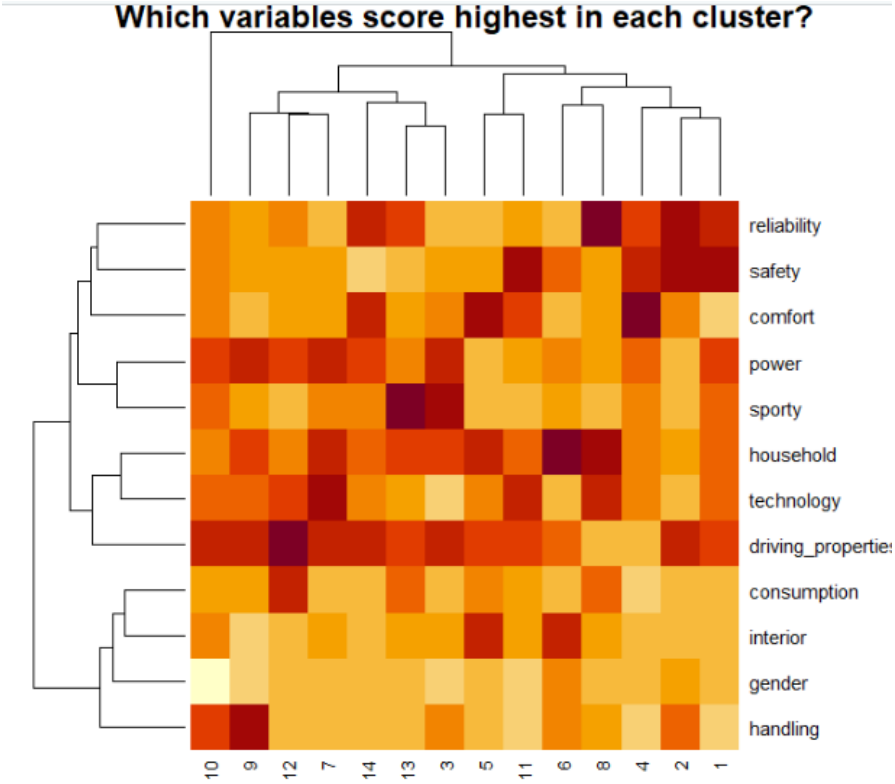
```
> table(assignments)
assignments
 1  2  3  4  5  6  7  8  9 10 11 12 13 14
64 64 93 32 74 43 53 39 90 49 47 51 54 40
>
```

Cutting down the tree to suggested 14 clusters

```
> clusterMeans
                        1         2         3       4         5         6         7         8         9        10        11        12        13    14
driving_properties 0.687500 0.734375 0.84946237 0.06250 0.62162162 0.20930233 0.88679245 0.00000000 0.81111111 1.0000000 0.7872340 0.90196078 0.53703704 0.975
interior           0.125000 0.140625 0.26881720 0.15625 0.68918919 0.32558140 0.26415094 0.10256410 0.06666667 0.6938776 0.3191489 0.00000000 0.14814815 0.050
technology         0.500000 0.125000 0.02150538 0.34375 0.40540541 0.04651163 1.00000000 0.61538462 0.57777778 0.7959184 0.9361702 0.56862745 0.14814815 0.350
comfort            0.062500 0.265625 0.39784946 0.93750 0.87837838 0.00000000 0.18867925 0.17948718 0.15555556 0.7551020 0.7659574 0.13725490 0.14814815 0.900
reliability        0.781250 0.750000 0.03225806 0.65625 0.14864865 0.00000000 0.11320755 0.87179487 0.23333333 0.6938776 0.2978723 0.25490196 0.61111111 0.925
handling           0.031250 0.453125 0.35483871 0.03125 0.12162162 0.16279070 0.00000000 0.10256410 0.94444444 0.9795918 0.0212766 0.05882353 0.11111111 0.075
power              0.687500 0.140625 0.73118280 0.46875 0.02702703 0.16279070 0.88679245 0.10256410 0.75555556 0.8979592 0.2340426 0.52941176 0.27777778 0.675
consumption        0.109375 0.125000 0.16129032 0.03125 0.32432432 0.00000000 0.09433962 0.46153846 0.22222222 0.6122449 0.2553191 0.70588235 0.40740741 0.125
sporty             0.593750 0.031250 0.92473118 0.40625 0.06756757 0.06976744 0.45283019 0.07692308 0.28888889 0.8571429 0.1276596 0.05882353 1.00000000 0.450
safety             0.921875 0.812500 0.23655914 0.75000 0.16216216 0.18604651 0.18867925 0.20512821 0.24444444 0.7551020 0.9787234 0.23529412 0.09259259 0.000
gender             0.093750 0.203125 0.02150538 0.15625 0.08108108 0.13953488 0.05660377 0.00000000 0.03333333 0.0000000 0.0000000 0.00000000 0.05555556 0.050
household          0.546875 0.218750 0.65591398 0.34375 0.74324324 0.48837209 0.84905660 0.74358974 0.67777778 0.6938776 0.6595745 0.33333333 0.53703704 0.650
```

A few observations

- Clusters 8, 10 – 12 are all male
- No one in cluster 8 rated driving_properties as important, while everyone in clusters 10 – and most people in clusters 7, 9, 12, and 14 – rated driving_properties as important
- Interestingly, cluster 10 has some of the highest scores across all of the features. This might be a cause for checking that data to see if those respondents just answered 1 for everything and might not have provided useful data.
- From the heatmap (below) we can see clusters 1, 2, and 4 really care about reliability; across all clusters driving_properties is often viewed as more important relative to other features; clusters 9 and 10 care more about handling – relative to other clusters – than other features; clusters 10, 9, 12, 7, and 14 care more about power – relative to other clusters – than other features



Problem 1h

Set up the k means clustering

```
##----- PROBLEM 1h -----##

# kmeans
NumberOfClusters = 14
set.seed(2407)
km = kmeans(data, iter.max=100, NumberOfClusters)
```

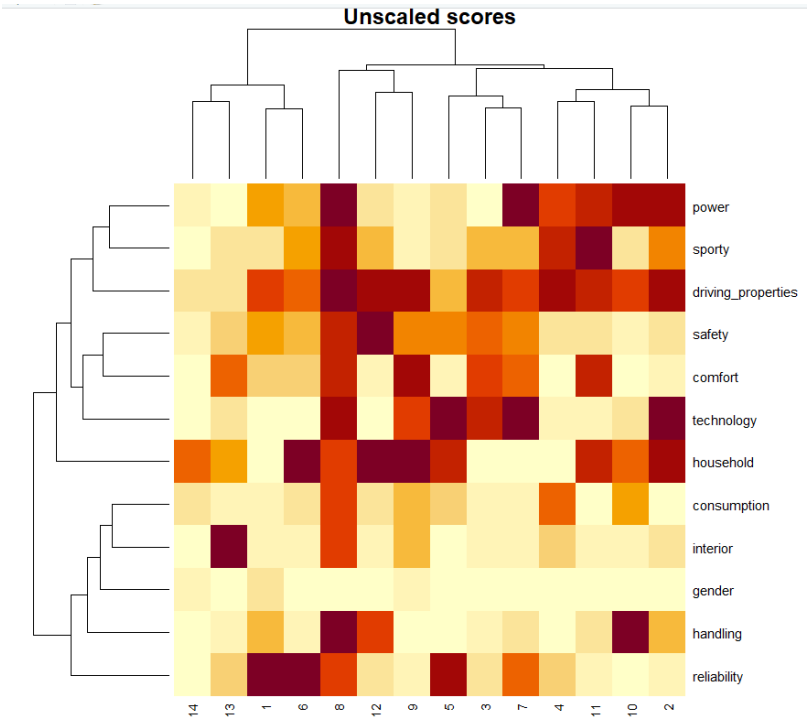
Using km.size, we get the breakdown of the number of data points in each cluster. The below numbers are ordered from cluster 1 – 14 accordingly

```
> km.size
 [1] 65 80 42 50 57 73 52 45 66 54 59 43 59 48
>
```

Column means for all clusters

```
> clusterMeans
                        1      2          3    4          5          6          7          8          9         10         11         12         13         14
driving_properties 0.73846154 0.8750 0.76190476 0.84 0.36842105 0.61643836 0.73076923 0.9777778 0.89393939 0.74074074 0.77966102 0.90697674 0.23728814 0.18750000
interior           0.10769231 0.1750 0.14285714 0.26 0.05263158 0.13698630 0.11538462 0.7111111 0.37878788 0.12962963 0.11864407 0.13953488 1.00000000 0.00000000
technology         0.03076923 1.0000 0.80952381 0.12 1.00000000 0.00000000 0.94230769 0.8666667 0.69696970 0.22222222 0.10169492 0.06976744 0.16949153 0.08333333
comfort            0.32307692 0.1000 0.69047619 0.08 0.10526316 0.27397260 0.61538462 0.8000000 0.90909091 0.07407407 0.76271186 0.09302326 0.62711864 0.04166667
reliability        1.00000000 0.1000 0.19047619 0.26 0.87719298 1.00000000 0.65384615 0.6888889 0.10606061 0.07407407 0.11864407 0.20930233 0.27118644 0.00000000
handling           0.35384615 0.3500 0.16666667 0.02 0.07017544 0.13698630 0.21153846 0.9333333 0.06060606 0.92592593 0.16949153 0.72093023 0.10169492 0.08333333
power              0.43076923 0.9000 0.00000000 0.68 0.21052632 0.34246575 1.00000000 0.9555556 0.10606061 0.90740741 0.81355932 0.20930233 0.06779661 0.10416667
consumption        0.09230769 0.0750 0.11904762 0.64 0.28070175 0.24657534 0.15384615 0.7111111 0.36363636 0.50000000 0.05084746 0.18604651 0.15254237 0.18750000
sporty             0.16923077 0.5625 0.35714286 0.80 0.24561404 0.46575342 0.40384615 0.8444444 0.10606061 0.22222222 0.96610169 0.34883721 0.18644068 0.06250000
safety             0.47692308 0.1750 0.64285714 0.18 0.56140351 0.35616438 0.57692308 0.7777778 0.51515152 0.09259259 0.22033898 0.93023256 0.27118644 0.10416667
gender             0.16923077 0.0625 0.07142857 0.06 0.00000000 0.06849315 0.03846154 0.0000000 0.09090909 0.03703704 0.03389831 0.02325581 0.03389831 0.14583333
household          0.00000000 0.8625 0.00000000 0.06 0.77192982 1.00000000 0.05769231 0.6888889 1.00000000 0.61111111 0.81355932 0.95348837 0.49152542 0.60416667
```

Unscaled heatmap



A few takeaways

- Similar to problem 1g, we see a specific set of clusters with high importance values for driving_properties, sporty, and power
- Additionally, similar to problem 1g there is a cluster (cluster 8) that has high values for every feature. Again this raises alarm bells because it could be a set of respondents that just fill out the survey without taking time to think through their answers. We might want to consider removing this set of respondents.
- Interestingly, in this clustering we *also* see a set of respondents that have low scores for all the values. It does not make intuitive sense that they do not see much importance in any of the features. Perhaps this is one group we *also* consider removing, assuming they just quickly filled out the survey.
- We would want to target cluster 13 with messaging about the interior and comfort and potentially design of the car given their high scores specifically for those features.

Problem 1i

I will frame the report from an advertising lens, using k-means clustering results.

With advertising, it is difficult to be all things to all people. There are certain costs involved with designing the artwork and messaging for an ad and executing on a specific strategy through channels such as TV, radio, podcasts, and social media. Therefore, I recommend that an ad campaign is developed based on a select few features that seem to resonate with a broad number of clusters (and in turn a broad number of customers). Looking at the unscaled heatmap in problem 1h, we find that power, sporty, driving_properties, and safety score high on importance across many of the clusters. Therefore, I

would recommend we develop an advertising campaign with messaging and videos that push those specific features of the car.

From a demographics standpoint, we can assume the customer of this car (potentially a German sports car) is mostly male. We assume that because most of the respondents to the survey were male (think it is safe to say that this result was not a matter of poor surveying techniques). So as an advertiser, I would recommend we go through channels with largely male audiences such as sporting events, sports sections of the newspaper, stadiums, and sports podcasts.