HW11

Rhea Toves

11/17/2021

Problem 1 is associated with your personal dataset project. Submit your work for all problems in a single .pdf compiled with knitr. The Homework Data folder has a starting .Rmd template with spaces for you to fill in your answers to the questions.

Problem 1 - This question is related to your personal data set project. The goal for the end of this week is to put together an initial visualization of your data and to perform one piece of analysis on your data.

(A) Use your dataset (possibly extended from what you originally submitted week 9) to create a visualization that summarizes an important property of the data. You should follow the principles of good visualization design that we discussed in week 4 and make sure to include a title, axis labels, and any other necessary components.

```
##                    title          artist                top.genre year
## 1         Blinding Lights      The Weeknd canadian contemporary r&b 2020
## 2       Watermelon Sugar    Harry Styles                      pop 2019
## 3 Mood (feat. iann dior)        24kGoldn                 cali rap 2021
## 4      Someone You Loved   Lewis Capaldi                      pop 2019
## 5                Perfect      Ed Sheeran                      pop 2017
## 6               Believer Imagine Dragons               modern rock 2017
##   beats.per.minute energy danceability loudness.dB liveness valance length
## 1              171     73           51          -6        9      33    200
## 2               95     82           55          -4       34      56    174
## 3               91     72           70          -4       32      73    141
## 4              110     41           50          -6       11      45    182
## 5               95     45           60          -6       11      17    263
## 6              125     78           78          -4        8      67    204
##   acousticness speechiness popularity
## 1            0           6         91
## 2           12           5         88
## 3           17           4         88
## 4           75           3         86
## 5           16           2         86
## 6            6          13         86
```

```
##
## Attaching package: 'dplyr'
```

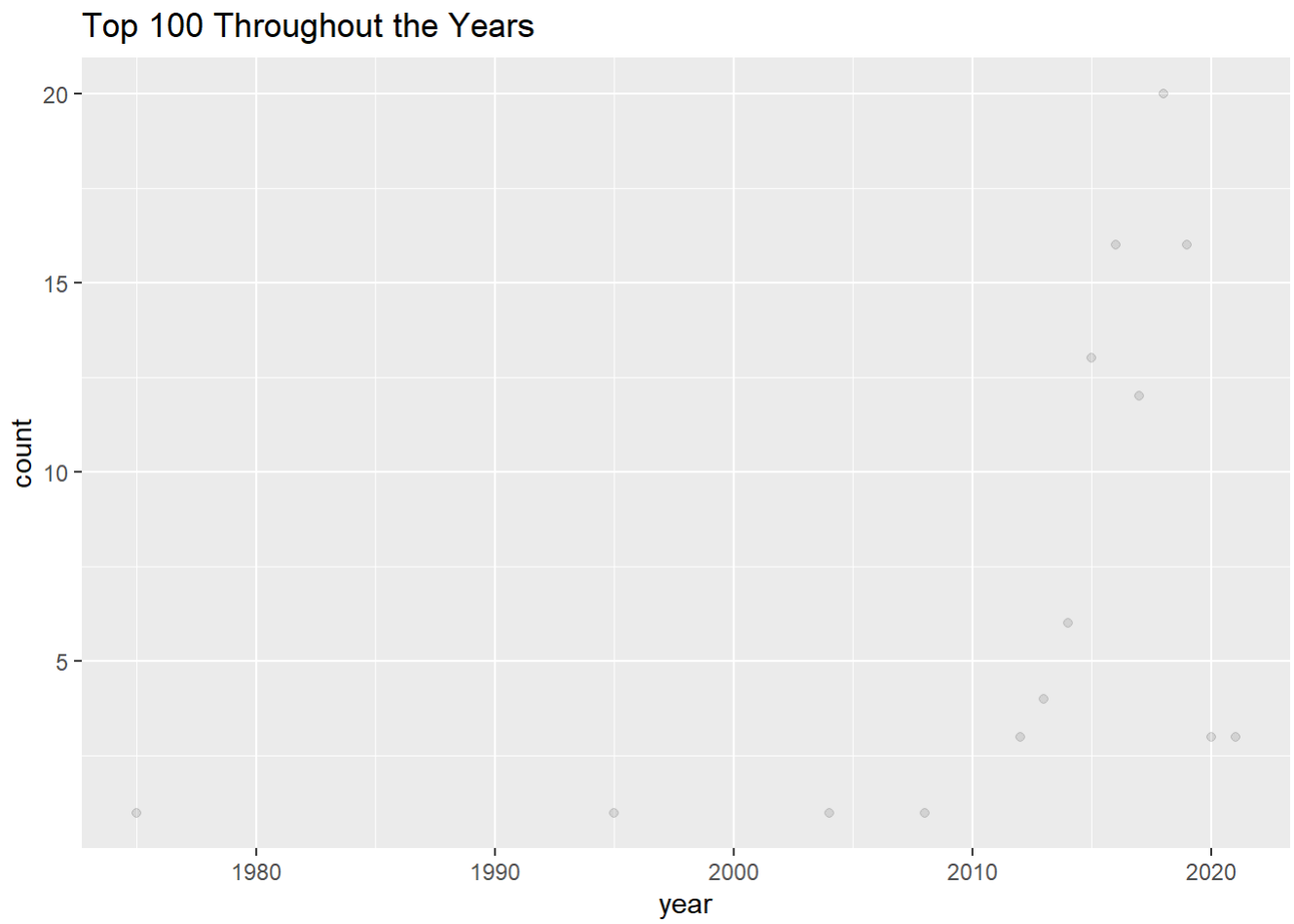
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

(B) Upload your figure to your github repository for the project and incorporate it into the readme file, along with a description (caption) of the plot that describes its contents.

(C) Choose one of the analytical techniques that we have discussed in class so far (either from the EDA section or the (un)supervised learning techniques) and apply it to your data. Explain in a brief paragraph why you chose this technique and what you learned about your data by performing this analysis (you don't have to upload this to github). This can be anything from a simple EDA technique like identifying outliers, making a box plot of a column, evaluating the summary statistics (Tukey's 5 numbers), to one of the regression models from supervised learning, or one of the clustering methods from unsupervised learning. This is not an exhaustive list of possibilities and as with the previous components doesn't need to be long or drawn out.

```
##                    title          artist                top.genre year
## 1         Blinding Lights      The Weeknd canadian contemporary r&b 2020
## 2        Watermelon Sugar    Harry Styles                      pop 2019
## 3 Mood (feat. iann dior)        24kGoldn                 cali rap 2021
## 4      Someone You Loved   Lewis Capaldi                      pop 2019
## 5                 Perfect      Ed Sheeran                      pop 2017
## 6               Believer Imagine Dragons              modern rock 2017
##   beats.per.minute energy danceability loudness.dB liveness valance length
## 1              171     73           51          -6        9      33    200
## 2               95     82           55          -4       34      56    174
## 3               91     72           70          -4       32      73    141
## 4              110     41           50          -6       11      45    182
## 5               95     45           60          -6       11      17    263
## 6              125     78           78          -4        8      67    204
##   acousticness speechiness popularity
## 1            0           6         91
## 2           12           5         88
## 3           17           4         88
## 4           75           3         86
## 5           16           2         86
## 6            6          13         86
```



For this question, I wanted to discover more information about the years the top 100 songs were created in. The EDA technique I focused on was identifying outliers. The scatterplot does a great job in pointing out the outliers, you can see that songs from the late 90's and early 2000's are outliers compared to the rest of the data. From the years 2015-2019, the top 100 songs fall more into this category of years. This was an interesting discovery.

Problem 2 - In your own words, please write brief answers to the following:

(A) What is the curse of dimensionality?

- When dimensionality of the features space increases and the number of configurations grow exponentially.

(B) What is the difference between MDS and PCA?

- MDS is the process in which we convert distances to low-dimensional embedding. PCA is the process in which we transform our data to represent the largest amount of variance with the fewest amount of data sets.

(C) Give an example of a dataset for which dimension reduction would be a useful first step.

- A dataset dealing with every name in the whole world would need to have dimension reduction.

(D) What is a dendrogram and what type of clustering method is it used to represent?

- A dendrogram is a graph that shows the clusters of data and the type of clustering method used is hierarchical clustering.

(E) What is the difference between supervised and unsupervised learning?

- Supervised learning is the process in which a data scientist would separate data into two types of problems: classification and regression. Unsupervised learning is the process in which a data scientist would use this

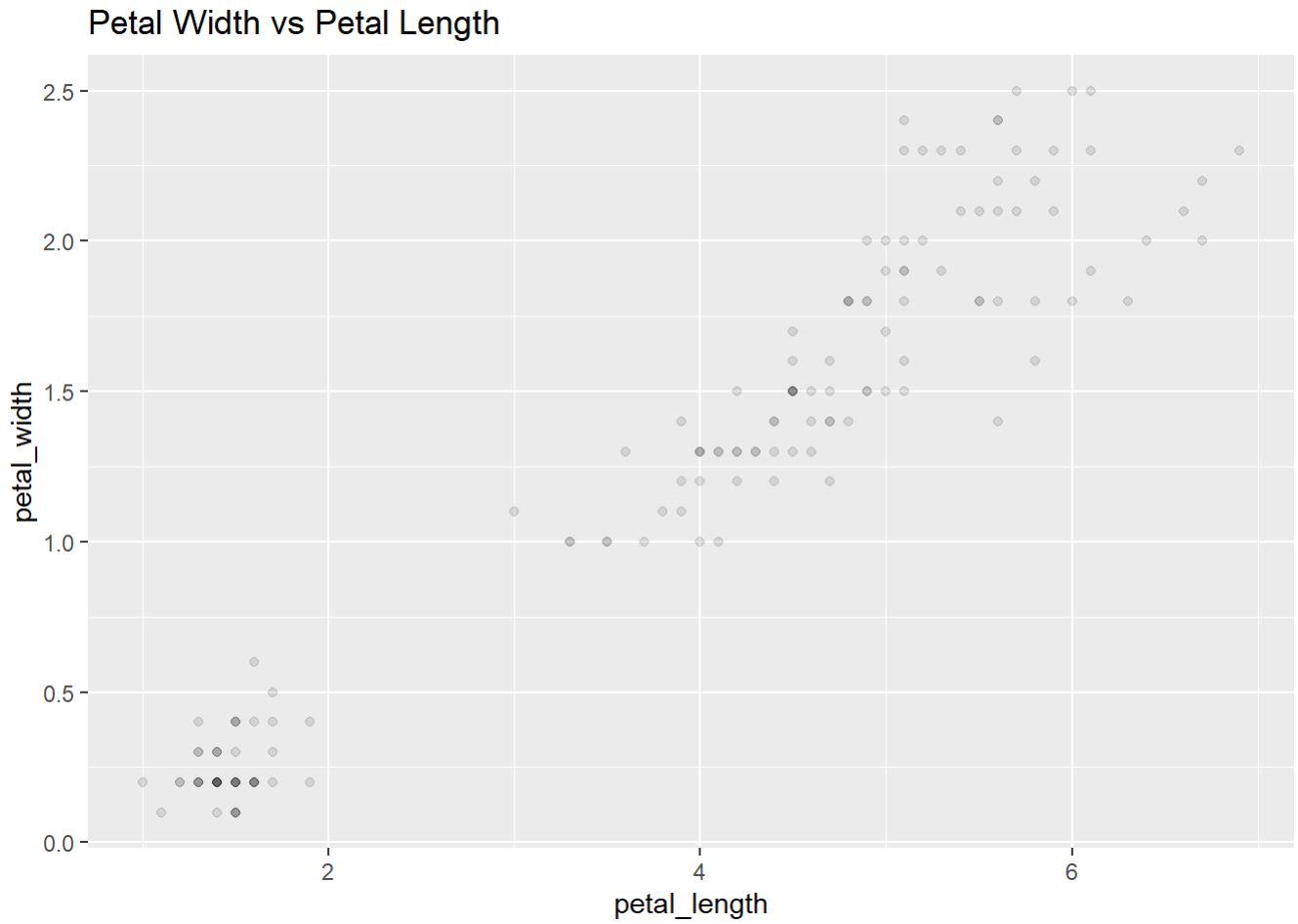
during the process of clustering, dimensionality, and association.

Problem 3 - This problem checks your understanding of PCA and K means clustering.

(A) Load the iris data as a dataframe. This is a classic dataset (originally published in 1936) that is frequently used as an initial set of test data. It consists of four numerical columns reporting the length and width of the sepal and petals of 150 iris plants and a final column reporting the specific subspecies. The version in the week 11 data folder also has an additional column with numerical column representing the subspecies.

```
##   sepal_length sepal_width petal_length petal_width   type type_numeric
## 1          5.1         3.5          1.4         0.2 setosa            0
## 2          4.9         3.0          1.4         0.2 setosa            0
## 3          4.7         3.2          1.3         0.2 setosa            0
## 4          4.6         3.1          1.5         0.2 setosa            0
## 5          5.0         3.6          1.4         0.2 setosa            0
## 6          5.4         3.9          1.7         0.4 setosa            0
```

(B) Make a scatterplot of petal width vs. petal length colored by the subspecies.



(C) Use R to perform PCA on the data with the four numerical columns as inputs and make a scatter plot of the top two principal components colored by subspecies.

```
## Importance of components:
##                           PC1    PC2     PC3     PC4
## Standard deviation     1.7061 0.9598 0.38387 0.14355
## Proportion of Variance 0.7277 0.2303 0.03684 0.00515
## Cumulative Proportion  0.7277 0.9580 0.99485 1.00000
```

(D) What proportion of the variance is explained by these two components?

- Across the two columns were explaining the 95% of the variance is across all the six columns from the original input.

(E) What are the loadings for each of the original numerical columns?

```
## [1] "sdev"     "rotation" "center"   "scale"    "x"
```

```
##         PC1         PC2         PC3         PC4
## -2.25698063 -0.50401540  0.12153619  0.02299628
```

(F) Apply K means clustering to the four numeric columns with three clusters.

```
## $cluster
##   [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
##  [38] 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
##  [75] 3 3 3 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 3 1 1 1 1 3 1 1 1 1
## [112] 1 1 3 3 1 1 1 1 3 1 3 1 3 1 1 3 3 1 1 1 1 3 1 1 1 1 3 1 1 1 3 1 1 1 3 1
## [149] 1 3
##
## $centers
##   sepal_length sepal_width petal_length petal_width
## 1     6.850000    3.073684     5.742105    2.071053
## 2     5.006000    3.418000     1.464000    0.244000
## 3     5.901613    2.748387     4.393548    1.433871
##
## $totss
## [1] 680.8244
##
## $withinss
## [1] 23.87947 15.24040 39.82097
##
## $tot.withinss
## [1] 78.94084
##
## $betweenss
## [1] 601.8836
```

(G) Apply K means clustering to the two principal components with three clusters.

```
## $cluster
##   [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [38] 1 1 1 1 1 1 1 1 1 1 1 1 1 3 3 3 2 3 2 3 2 3 2 2 2 2 2 2 3 2 2 2 2 2 2 2
##  [75] 3 3 3 3 2 2 2 2 2 2 2 2 3 2 2 2 2 2 2 2 2 2 2 2 2 3 2 3 3 3 3 2 3 3 3 3
## [112] 3 3 2 2 3 3 3 3 2 3 2 3 2 3 3 2 2 3 3 3 3 3 2 2 3 3 3 2 3 3 3 2 3 3 3 2 3
## [149] 3 2
##
## $centers
##   sepal_length sepal_width
## 1     5.006000    3.418000
## 2     5.773585    2.692453
## 3     6.812766    3.074468
##
## $totss
## [1] 130.1809
##
## $withinss
## [1] 13.2020 11.3000 12.6217
##
## $tot.withinss
## [1] 37.1237
##
## $betweenss
## [1] 93.05723
```

(H) Which of the two K means clusterings is more accurate at predicting the subspecies correctly?

- The first k means clustering is more accurate at predicting the subspecies correctly.

Problem 4 - Use the following exploratory data analysis projects about candy preferences: https://fivethirtyeight.com/videos/the-ultimate-halloween-candy-power-ranking/ (https://fivethirtyeight.com/videos/the-ultimate-halloween-candy-power-ranking/) to provide brief responses to the following:

(A) What research questions are the articles trying to answer? Where does your favorite candy rank in their list?

- The authors stated they were curious as to what Halloween candy people preferred and performed research based on this question. My favorite candy, Milk Duds, placed 32/86.

(B) Can you tell where the underlying data came from?

- The underlying data came from a survey with 8,371 different IP addresses who voted on 269k randomly generated match-ups.

(C) What are some other ways you could use to determine which candy is best?

- You could have people insert (type) their favorite candy so they are not limited on the options given to them in the survey.

(D) What type of regression is used in the candy power ranking article? How does the interpretation of the coefficients match up with what we discussed in class?

- The specific regression was not mentioned in the article. But the interpretation of the coefficients match up with what we discussed in class.

(E) What do you think about the visualizations and examples in these pieces?

- The visualizations were very helpful in putting these statistics in perspective.