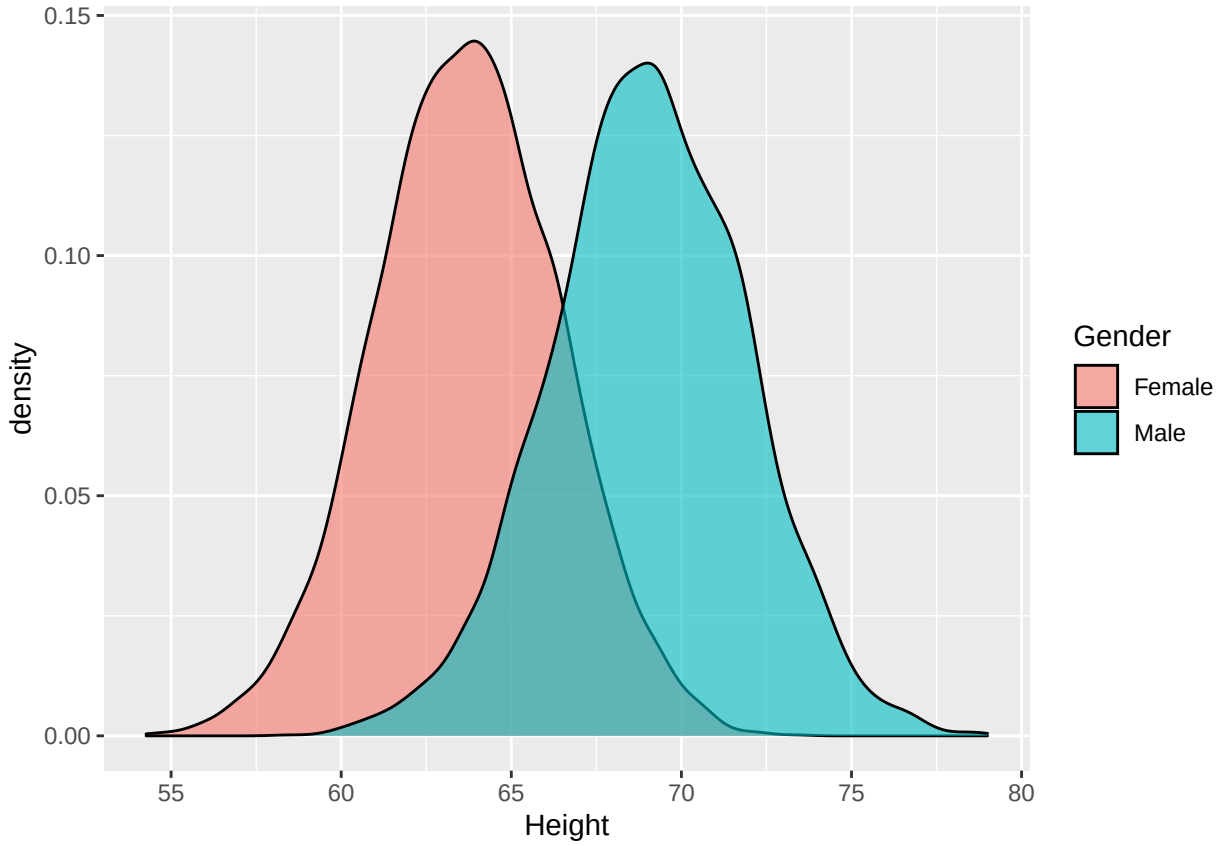
HW5

Peyton Kuhlers

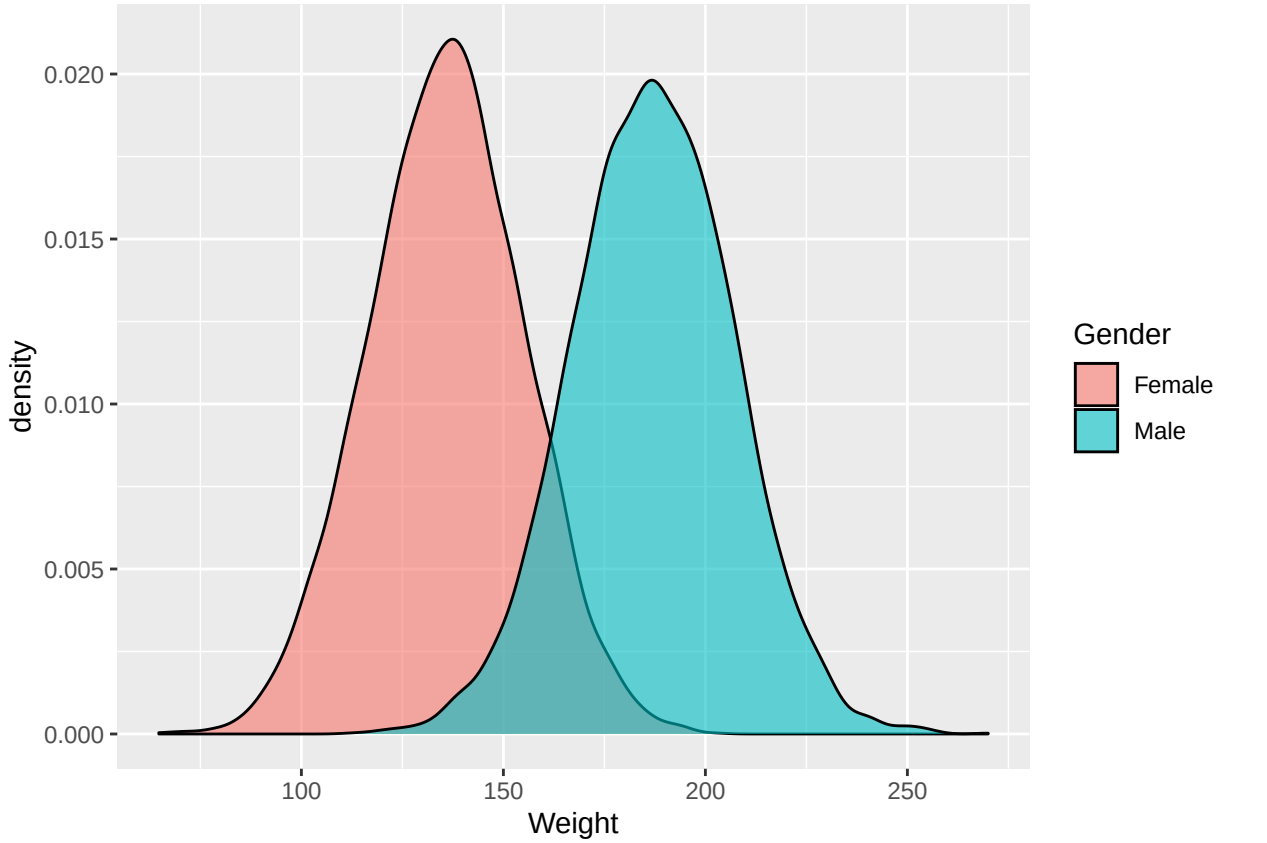
10/29/2020

Question 1

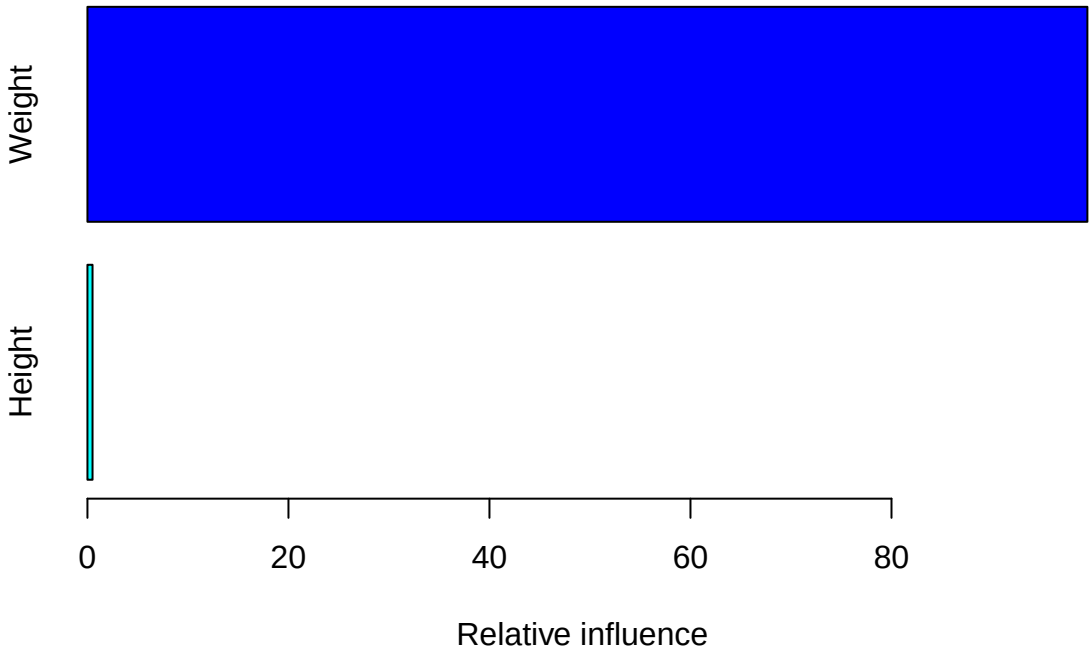
```
suppressMessages(library(gbm))
suppressMessages(library(tidyverse))
suppressMessages(library(MLmetrics))
suppressMessages(library(caret))

dat <- read.csv("datasets_weight-height.csv")
dat$Gender_num <- ifelse(dat$Gender == "Male", 0, 1)

ggplot(data = dat, aes(x = Height, fill = Gender)) +
  geom_density(alpha = 0.6)
```




```
ggplot(data = dat, aes(x = Weight, fill = Gender)) +
  geom_density(alpha = 0.6)
```




```
mod <- gbm(formula = Gender_num ~ Height + Weight,
           distribution = "bernoulli",
           data = dat)
summary(mod)
```



```
##           var    rel.inf
## Weight Weight 99.4924084
## Height Height  0.5075916
```

Question 3

There appears to be a duplication in the dataset.

```
dat2 <- read.csv("datasets_charcters_stats.csv")
any(duplicated(dat2$Name))
```

```
## [1] TRUE
```

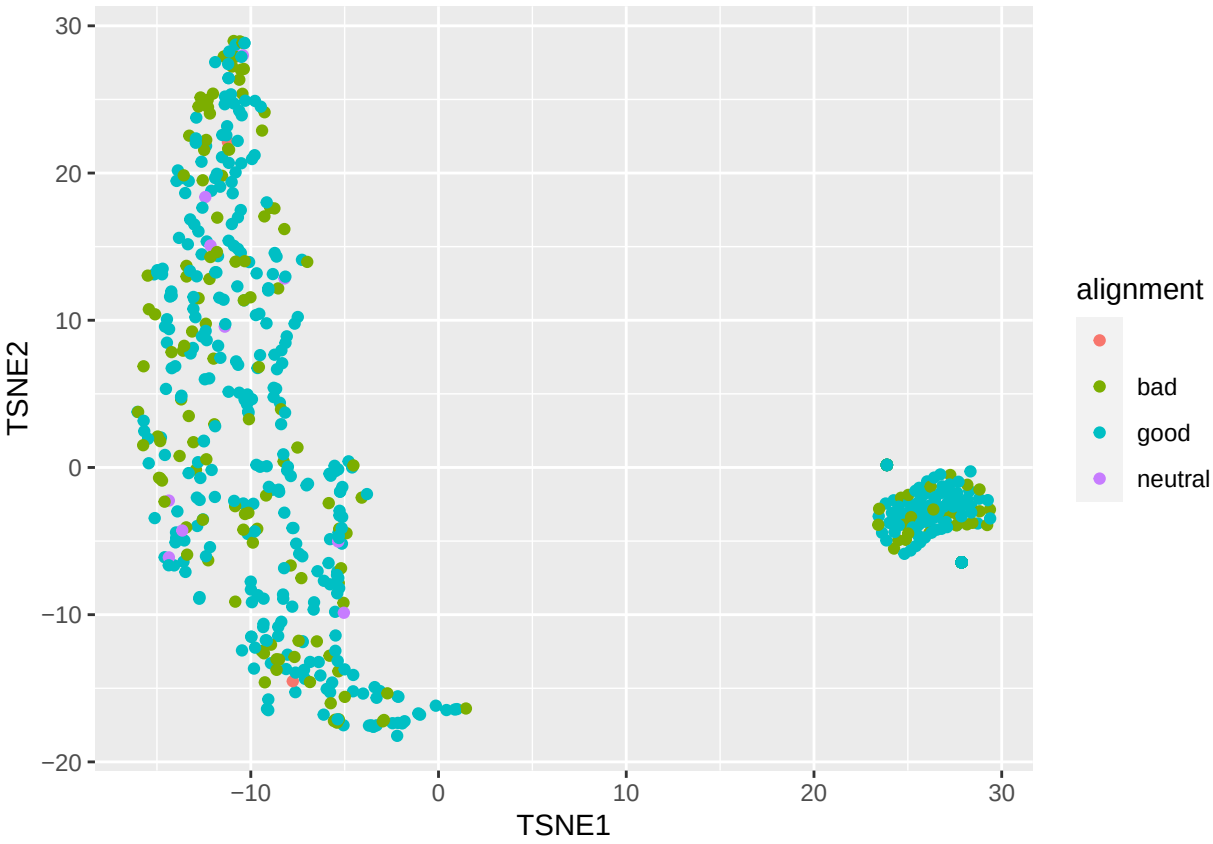
Here we will filter the duplication and run pca

```
pca <- dat2 %>%
  distinct(Name, .keep_all = T) %>%
  column_to_rownames(var = "Name") %>%
  select(c(-Alignment,-Total)) %>%
  prcomp(.)
```

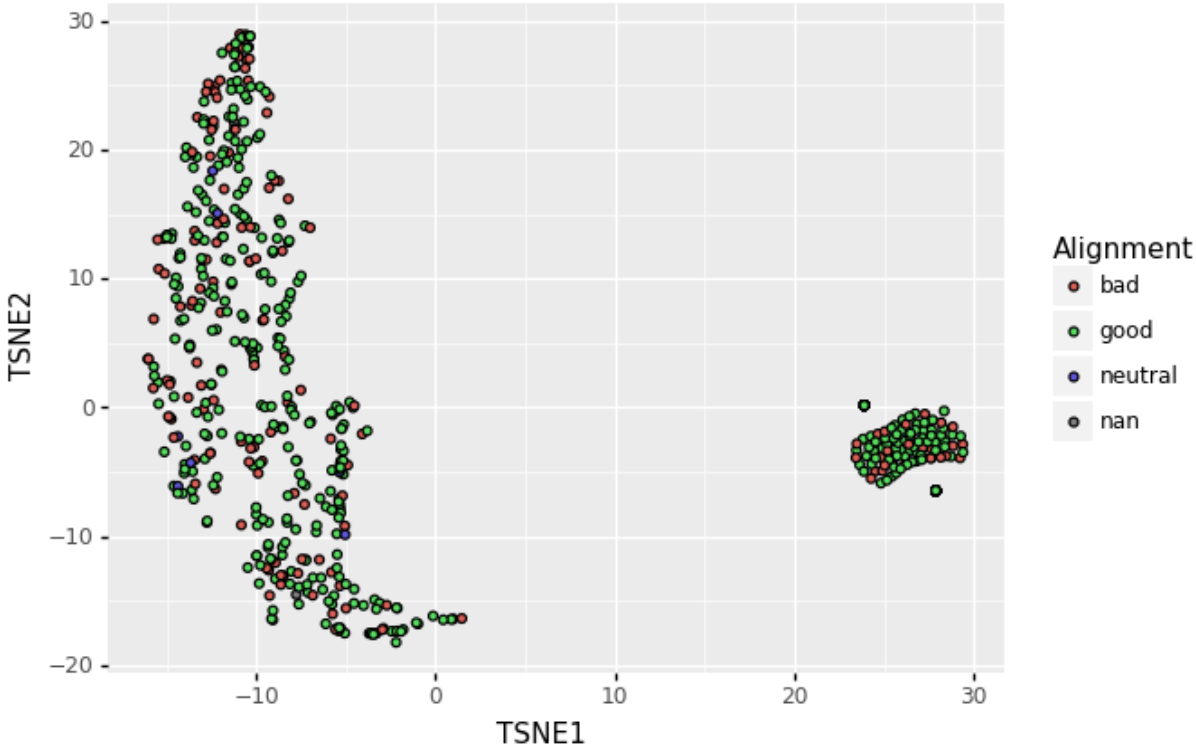
Question 3

We will now plot the TSNE data from sklearn.

```
tsne <- read.csv("tsneOut.csv")
align <- data.frame(alignment = dat2$Alignment,
                    TSNE1 = tsne$TSNE1,
                    TSNE2 = tsne$TSNE2)

ggplot(data = align, aes(x = TSNE1, y = TSNE2, col = alignment)) +
  geom_point()
```



And the plot produced by python plotnine:



Question 6

We like to use k-fold cross validation because it provides an estimate of test error and mimics the random process of data collection. That is, when we collected the data we could have gotten many potential values – so by training on many different folds of the data we get a better pictures of what happens if we had a different random selection. Additionally, by using k-folds we can get a variance for our estimate of test error as well. It's unwise to just report one training set error because the training error will necessarily be better than test error – by virtue of the fact that the model "knows" the data it was trained on.

Question 7

Recursive feature elimination is the process of starting will a model (can be a full or null model) and recursively adding or subtracting features from it until we get a model that is sufficiently large to have good predictive quality, but not so large that it overfits the data.