# HW_6_Geary

Marion Geary

2/19/2022

## Exercise 1

```
library(tidymodels)
library(gapminder)
data("gapminder")
```

## Exercise 2

```
gapminder_wide <- gapminder %>% pivot_wider(id_cols = c(country, continent), names_from = year, values_
# :)
```

## Exercise 3

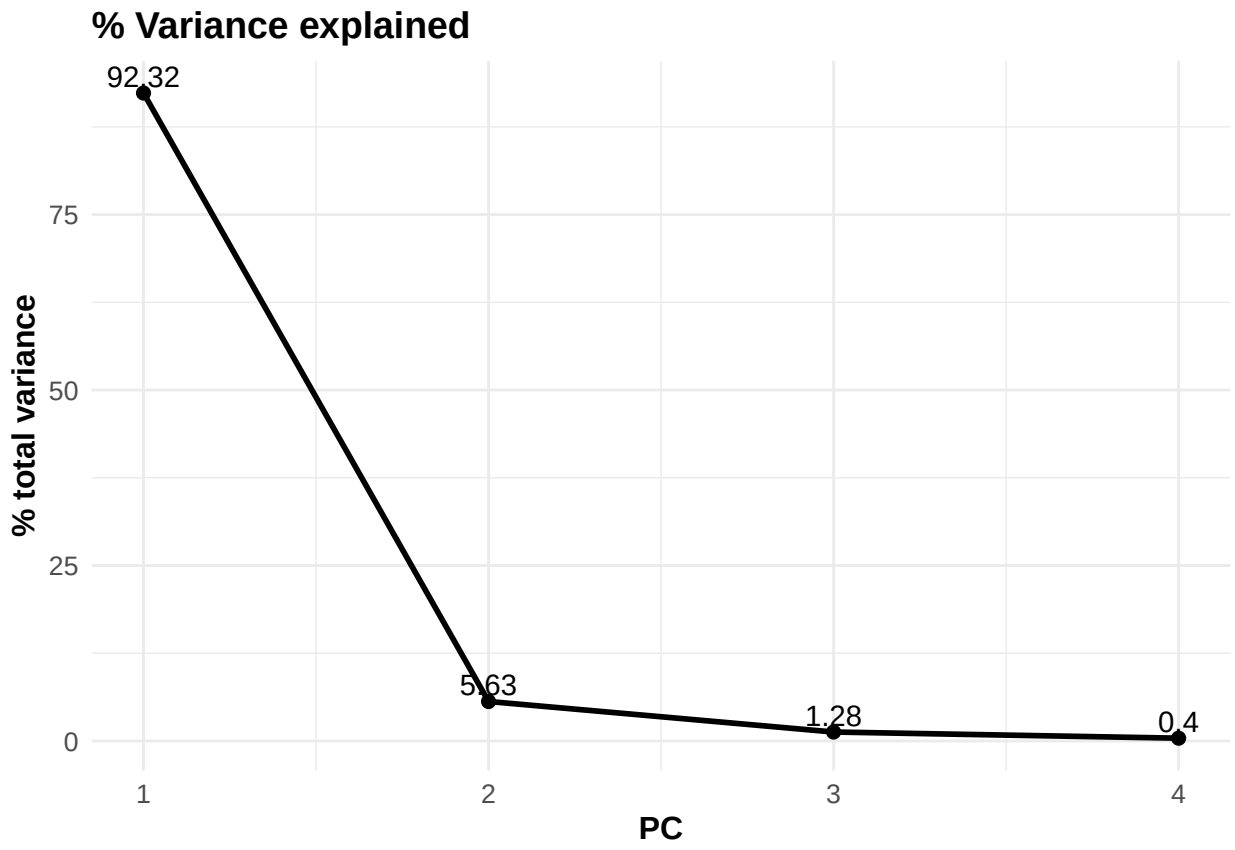
```
gapminder_recipe <- recipe(~., data = gapminder_wide) %>%
  step_normalize(all_numeric()) %>%
  step_pca(all_numeric(), num_comp = 4)

gapminder_prep <- prep(gapminder_recipe)
```

## Exercise 4

```
gap_vars <- tidy(gapminder_prep, 2, type = "variance")
gap_vars %>%
  filter(terms ==  "percent variance") %>%
  filter(component == c(1, 2, 3, 4)) %>%
ggplot(aes(x = component, y = value)) +
  geom_point(size = 2) +
  geom_line(size = 1) +
  scale_x_continuous(breaks = 1:4) +
  labs(title = "% Variance explained",
       y = "% total variance",
       x = "PC") +
  geom_text(aes(label = round(value, 2)), vjust = -0.3, size = 4) +
  theme_minimal() +
```

```
  theme(axis.title = element_text(face = "bold", size = 12),
        axis.text = element_text(size = 10),
        plot.title = element_text(size = 14, face = "bold"))
```

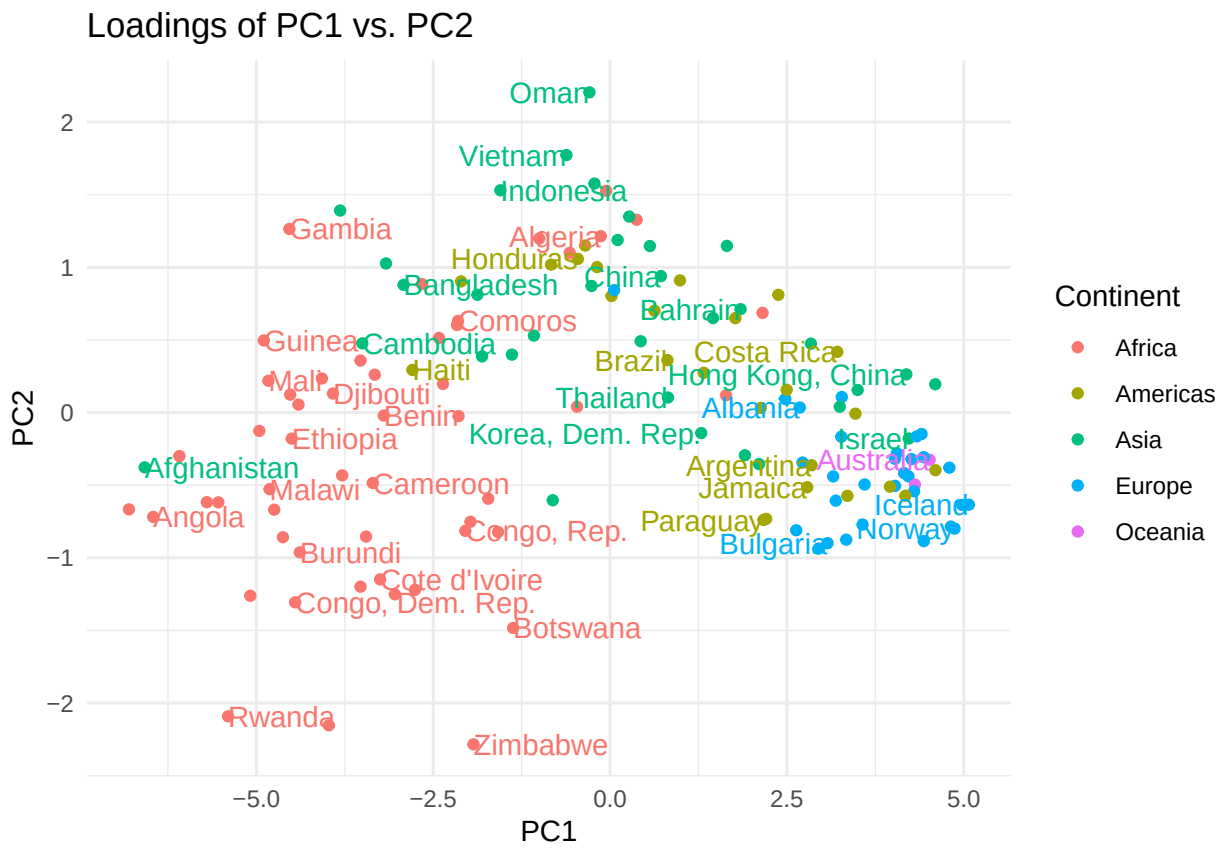



From the scree plot, we can see that a huge amount of the variance is explained by `PC1`, about 92%. The graph then quickly levels off, with the second principal component explaining about 5% of the variance and the next `PC`s explaining less. This graph shows that we can explain a vast majority of the variance with just one or two `PC`s.

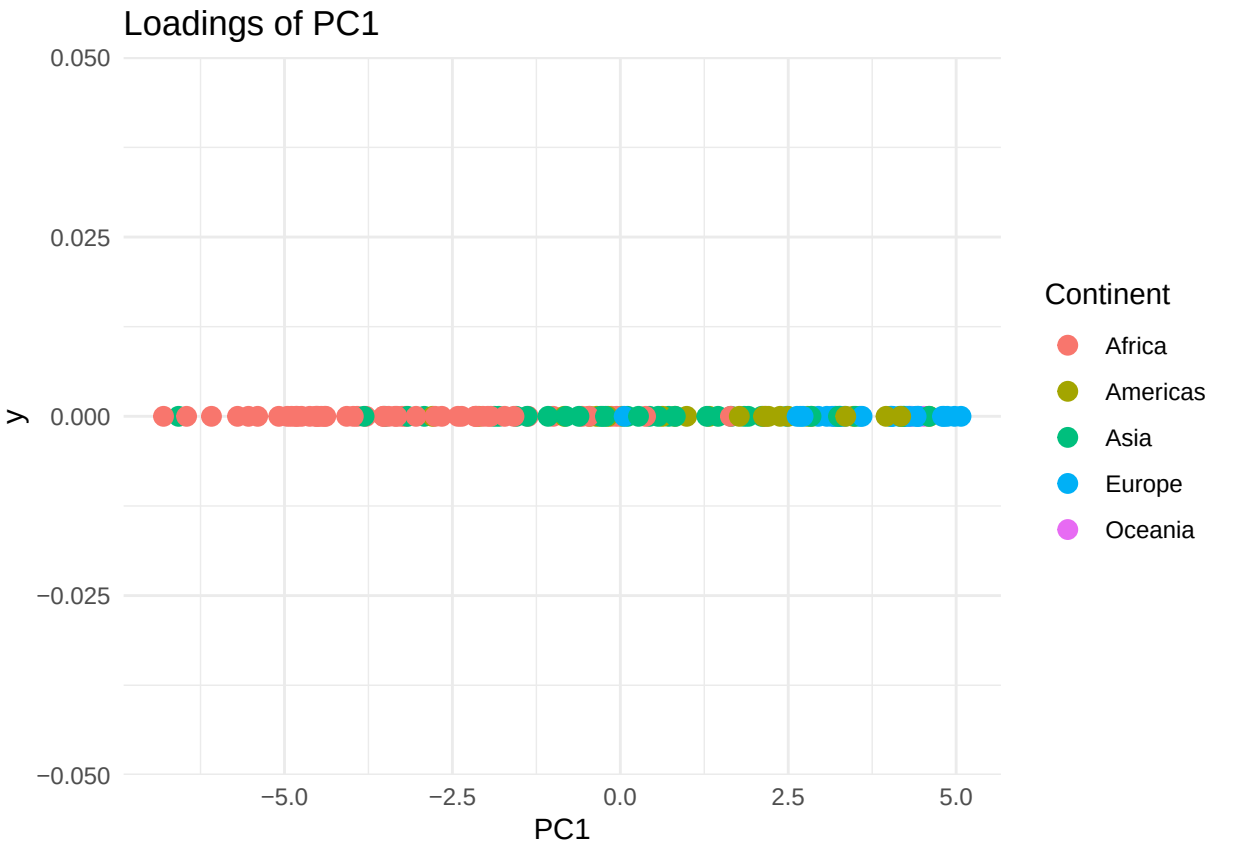
## Exercise 5

```
gapminder_juice <- juice(gapminder_prep)
# juice(prep) == bake(prep, new_data = prep_data)
# juicing is baking with whatever data you used to prep (probably training)

gapminder_juice %>% ggplot(mapping = aes(x = PC1, y = PC2, label = country, color = continent)) +
  geom_point() +
  geom_text(check_overlap = TRUE, hjust = "inward", show.legend = F) +
  labs(title = "Loadings of PC1 vs. PC2", color = "Continent") + theme_minimal()
```




```
# extra graph for analysis
gapminder_juice %>% ggplot(mapping = aes(x = PC1, y = 0, color = continent)) +
  geom_point(size = 3) +
  labs(title = "Loadings of PC1", color = "Continent") +
  theme_minimal()
```




From looking at these graphs, we can see clusters by continent moving left to right from `Africa` to `Asia` to `Americas` to `Europe`. We see a significant amount of separability based on the x axis alone, meaning that we probably could use just `PC1` for an effective analysis. Looking at the second graph, we see that we could identify clusters fairly well based on `PC1` alone.

I would probably use both `PC1` and `PC2` in my own modeling, because the vertical axis elucidates the clusters even more, and one more variable doesn't make a huge difference. For example, without the vertical axis, the `Americas` cluster would be easily mixed with `Asia` and `Europe`, so `PC2` provides a little more separability. Some clustering could get lost without this second variable, so while I don't think it's entirely necessary, I would probably keep it.