Project 3

Xiangrui Pan

XP742

This is the dataset used in this project:

```
# Get the Data

volcano <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2
```

Part 1

Question: How does elevation level of volcano compared in ascending years after 1900 in the United States, Japan, and Chile?

To answer this question, I will create a summary table and one visualization. In the summary table, it will have four columns: country, the mean of elevation, year of the most occurrences, and number of eruptions.

Introduction: A volcano is a rupture in the crust of a planetary-mass object, such as Earth, that allows hot lava, volcanic ash, and gases to escape from a magma chamber below the surface. This dataset comes from The Smithsonian Institution and it collects all the major volcano activities around the world after 1883. The dataset contains total of 26 columns that provides the type of volcano, years of eruption, population nearby the volcano, location of volcano, and elevation of volcano. We only need to work with some of them for part 1. We will be working with four columns in part 1:

1) `last_eruption_year`: This column includes the years of volcano eruptions
2) `evidence_category`: The column shows type of evidence of volcano eruptions
3) `elevation`: The height of volcano
4) `country`: The countries shows confirmed volcano activities

Approach: Our approach to the question in part 1 is first to build a summary table that includes the sample data for United States, Japan, and Chile so that we may gain some intuitions about what the plot should look like. Notice that `last_eruption_year` is encoded as character, We will use `as.numeric()` to make it numeric. We also want to find mode (the most occurrences) in the table, so I write a function `getmode()` to do it. Although I could use `mfv()` function from `library("modeest")`, it requires installing a package so I did not want to do it. Next, we will visualize the height of volcano against years of eruptions by using scatter points and regression line. We can easily see the trend of the height of volcano changes with respect to the change in years. We also can use line plot as oppose to using scatter plot but since we do not know if there is linear relation between height of volcano and years of eruptions, so scatter plot is more useful in this case.

I will list the following functions to make the summary table

1) `filter()`: to filter out the three countries, confirmed observed eruptions(not all the eruptions were confirmed), pick the year after 1900, and the volcano which were above the surface.
2) `select()`: select four needed columns
3) `mutate()`: put three countries in order. From Asia to North America to South America by using `fct_relevel()`

4) `group_by()`: take `country` to further operation

5) `summarize()`: receive the column from `group_by()` and make new columns for the summary table

Then, there are other functions for the visualization part. I use the above functions for visualizing the plot, so I am not going to introduce them again because they are used for the same purposes.

1) `geom_point()`: to create scatter plot of height of volcano for each year

2) `geom_smooth()`: add regression line to the scatter plot to see the trend

3) `facet_wrap()`: separate each plot for each of three countries

4) `scale_x_continuous()`: refine x-axis

5) `scale_y_continuous()`: refine y-axis

6) `theme()`: allows us to manually adjust the features of axis text.

7) `theme_bw()`: set black/white to the background

Analysis:

```
v = as.numeric(volcano$last_eruption_year)
```

```
## Warning: NAs introduced by coercion
```

```
#create a function to calculate mode (the most occurrences) in eruption year
getmode = function(v){
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}

#create the summary table
summary_table1 = volcano%>%
  filter(country %in% c("United States","Japan","Chile"),
         evidence_category %in% c("Eruption Observed"),
         last_eruption_year != "Unknown",
         #change last_eruption_year to number
         as.numeric(last_eruption_year) > 1900,
         elevation > 0
         )%>%
  select(
         last_eruption_year, country,
         elevation, evidence_category
         )%>%
  mutate(
    country = fct_relevel(country, "Japan", "United States","Chile")
  )%>%

  group_by(
    Country = country
    )%>%
  summarize(
   #Eruption_Observed = sum(evidence_category =="Eruption Observed"),
   Most_Eruption_Year = getmode(last_eruption_year), #get the mode
   Mean_Elevation = mean(elevation), #get the mean of elevation
   #get the total number of volcano that been described as observed
   Total_Confirmed_Eruption = length(evidence_category)
  )
```

```
## Warning in mask$eval_all_filter(dots, env_filter): NAs introduced by coercion
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

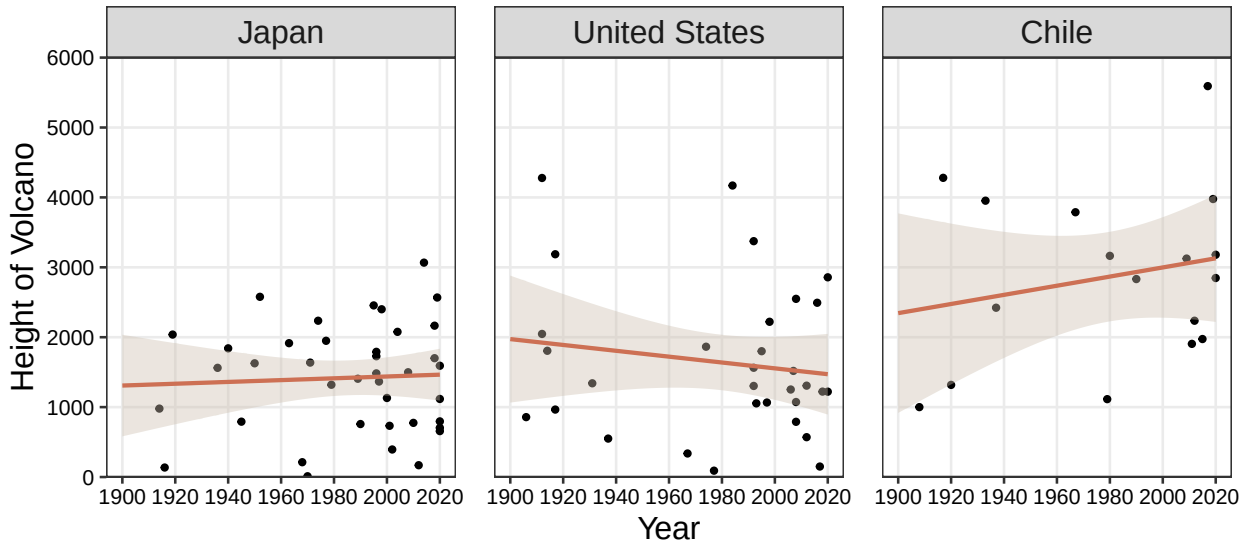
```
summary_table1
```

```
## # A tibble: 3 x 4
##   Country       Most_Eruption_Year Mean_Elevation Total_Confirmed_Eruption
##   <fct>         <chr>                       <dbl>                     <int>
## 1 Japan         2020                        1420.                        39
## 2 United States 1992                        1641.                        31
## 3 Chile         2020                        3047.                        18
```

```
#Visualization
 volcano%>%
  filter(
         country %in% c("United States","Japan","Chile"),
         evidence_category %in% c("Eruption Observed"),
         last_eruption_year != "Unknown",
         as.numeric(last_eruption_year) > 1900,
         elevation > 0
         )%>%

  select(
    last_eruption_year,
    country,
    elevation,
    evidence_category)%>%

  mutate(
    country = fct_relevel(
      country,
      "Japan", "United States","Chile"
      )
  )%>%

  group_by(
    country,
    last_eruption_year,
    elevation)%>%
  summarize(
    country,
    last_eruption_year,
    elevation
  )%>%

  ggplot(aes(
    x = as.numeric(last_eruption_year), #set years from string to number
    y = elevation
  ))+
  geom_point(size = 0.8) +
  geom_smooth(
    method = "lm",
    size = 0.75,
    fullrange = TRUE,
    color = "salmon3",
```

```
    fill = "antiquewhite3"
   )+
  facet_wrap(vars(country))+

   scale_x_continuous(
    name = "Year",
    limits = c(1900, 2020),
    breaks = seq(from = 1900, to = 2020, by = 20),
    labels = seq(from = 1900, to = 2020, by = 20),
    expand = c(0.05, 0.05))+

   scale_y_continuous(
    name = "Height of Volcano",
    limits = c(0, 6000),
    expand = c(0, 0)) +

    theme_bw(10)+

    theme(
      axis.title = element_text(size = 12),
      axis.text = element_text(color = "black", size = 7.5),
      panel.grid.minor = element_blank(),
      panel.spacing = unit(1, "lines"),
      strip.text.x = element_text(size = 12),
      aspect.ratio = 1.2
  )
```

```
## Warning in mask$eval_all_filter(dots, env_filter): NAs introduced by coercion
## `summarise()` regrouping output by 'country', 'last_eruption_year' (override with `.groups` argument)
## `geom_smooth()` using formula 'y ~ x'
## Warning: Removed 1 rows containing non-finite values (stat_smooth).
## Warning: Removed 1 rows containing missing values (geom_point).
```



Discussion: From the summary table, we can see that both Japan and Chile have reported to have most confirmed eruptions in 2020 and the US has most occurrences for active volcano in 1992. Moreover, Japan has total of 39 recorded, confirmed volcano eruptions, which is twice as active volcano eruptions as in Chile after 1900. Chile has the highest mean of height of volcano among these three countries.

We can gain more knowledge about the relationship between height of volcano and years of eruptions by looking at the regression line from our plots. In Japan, the regression line looks no slope and the variance of the height of volcano is smallest among the three. In Japan, Most of active volcano occurs after 1980. Also, most of active volcano occurs after 1980 in the United States since the points are more concentrated between 1980 and 2020. On the other hand, there is a clearly increasing trend for elevation of volcano against time in Chile and it has largest variance in elevations during time interval between 1900 and 2020. We can see that the highest volcano in Chile is around 5500 meters! Comparing the third plot with the first two plots, we can see Chile has higher rate of growing in elevation of volcano than Japan an US. The number of active volcano occurrences in Chile is much smaller than Japan and US. Since Japan and US are two industrialized-oriented countries and growing very fast after 1970. Maybe we can make further assumption about the occurrences of volcano has positive correlation with industrialization.

Part 2

Question: How can we perform dimension reduction method to do exploratory data analysis to investigate China and UK's volcano in the dataset?

Introduction: In part 2, we are interested in using less features to describe China and UK's volcano activities. As mentioned earlier, there are total of 26 features in this dataset. Some of them are less informative than the others, so we want to use more informative features to answer our question in part 2. To do that, I will perform Principal Component Analysis (PCA). I will first do a rotation plot for principal components 1 and

2, and the amount of variance explained by the various components. Next, I will make a scatter plot of PC 2 versus PC 1, color by these two countries. I will use all the features in the dataset before performing PCA except `volcano number`. The remaining 25 features, besides the four I described in part 1, are including geospatial data, type of volcano rock, population around volcano.

Approach: Our approach to part 2 is to perform PCA. PCA seeks a small number of dimensions that are as interesting as possible, where the concept of "interesting" is measured by the amount that the observations vary along each dimension. I could potentially use other dimension reduction methods like t-SNE and UMAP, but I do not know if the data looks spiral or not. Hence, PCA is safer choice in this case.

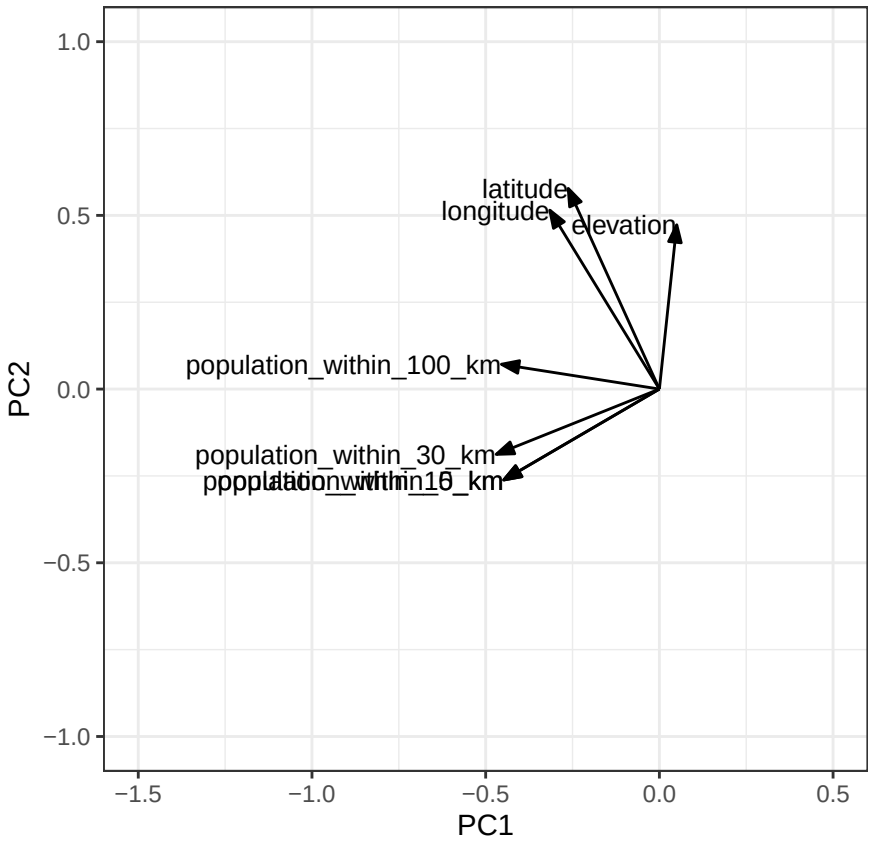
Before performing PCA on the dataset, I will drop the `volcano_number` from the dataset since it is just the index number of volcano, it is just like numeric name for each volcano. I will first run a rotation plot to see what features that explained by PC1 and PC2. Next, I will make a proportion variance explained (PVE) plot that shows us how much variation that each PC explains, so that we are able to know how much proportion of variation that PC1 and PC2 together explain in the model. Normally we just need the first two PCs because these two explain majority variations. Finally, we will make a scatter plot of volcano activities for China and UK to see how each point distributed in the plot and compare with the rotation plot to see which feature is relevant to the volcano activities. The following are the functions I used for my analysis:

1) `filter()`: to filter out China and United Kingdom
2) `where()`: find out numeric columns
3) `select()`: PCA only works for numeric values so we want to select numeric columns
4) `scale()`: scale to zero mean and unit variance
5) `prcomp()`: compute rotation matrix
6) `arrow()`: customized the arrow in rotation plot
7) `tidy()`: extract rotation matrix
8) `geom_text()`: customize the text in rotation plot
9) `coord_fixed()`: to fix the scale coordinate system
10) `geom_col()`: to show the proportion variance explain for each principal component

Analysis:

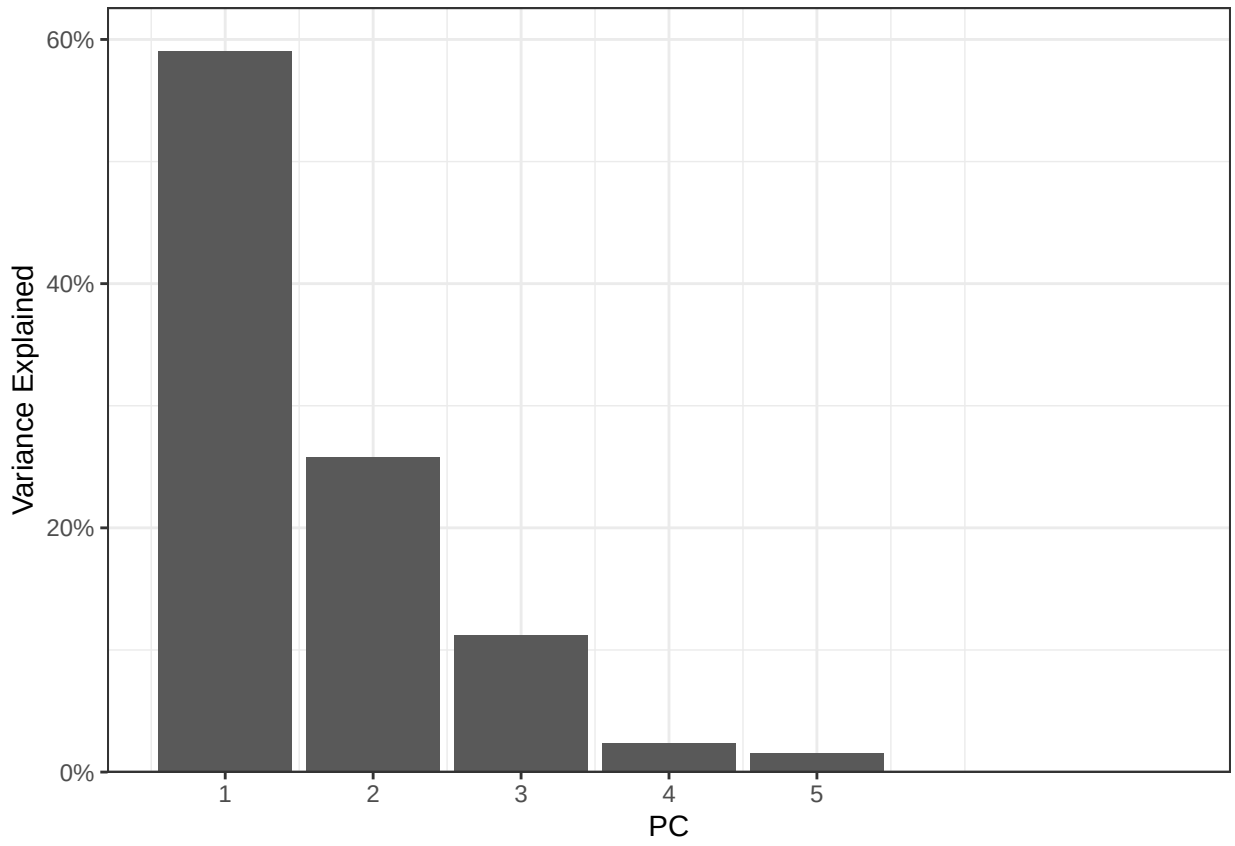
```
#remove the first column which is just volcano number, useless to our analysis
volcano = volcano[,-1]

volcano_update = volcano%>%
                     filter(
                       country%in%c(
                                    "China","United Kingdom"
                                    )
                     )

pca.fit = volcano_update %>%
  select(where(is.numeric)) %>% # retain only numeric columns
  scale() %>%                   # scale to zero mean and unit variance
  prcomp()




#customize the rotation arrow
arrow_style <- arrow(
```

```
  angle = 20, length = grid::unit(7, "pt"),
  ends = "first", type = "closed"
)

#rotation plot
pca.fit %>%
  # extract rotation matrix
  tidy(matrix = "rotation") %>%
  pivot_wider(
    names_from = "PC", values_from = "value",
    names_prefix = "PC"
  ) %>%
  ggplot(aes(PC1, PC2)) +
  geom_segment(
    xend = 0, yend = 0,
    arrow = arrow_style
  ) +
  geom_text(aes(label = column), hjust = 1,
            size = 3.5) +
  xlim(-1.5, 0.5) + ylim(-1, 1) +
  coord_fixed()+
  theme_bw()
```

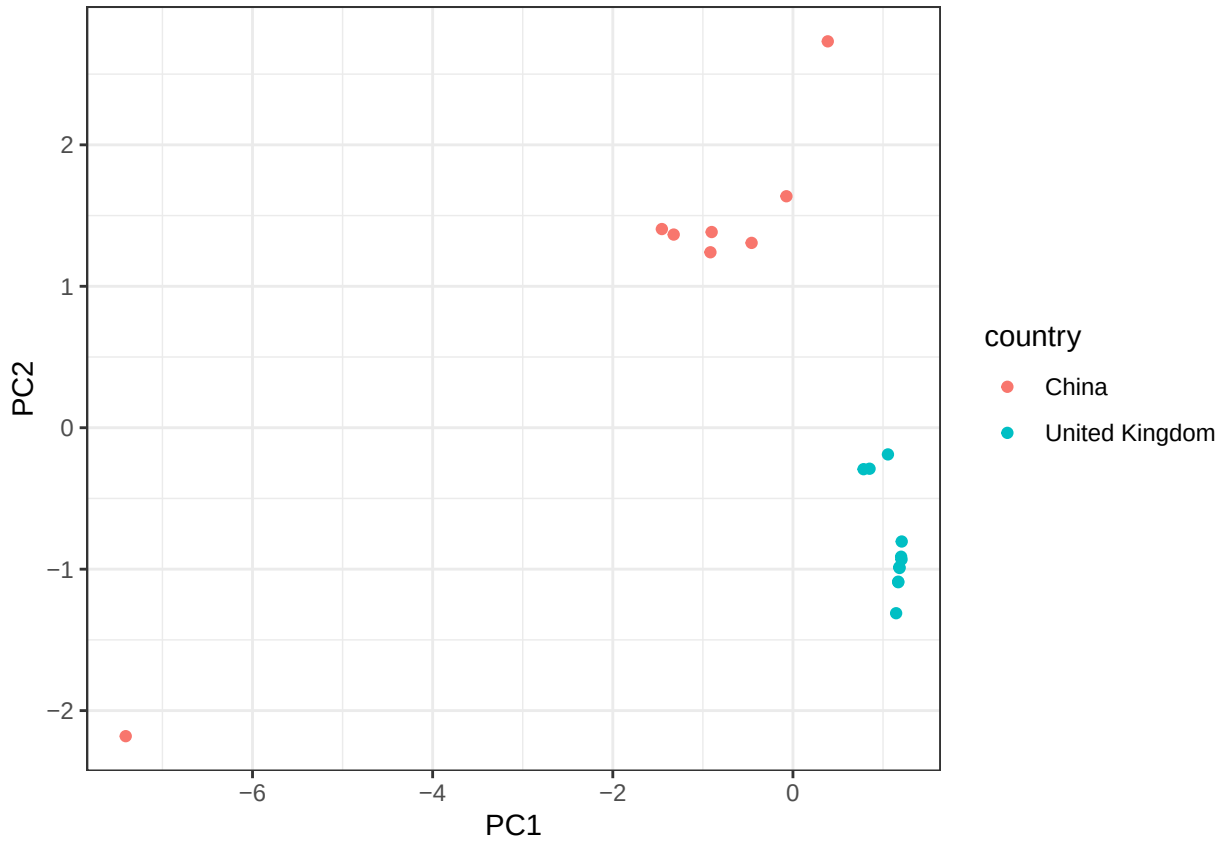



```
#PVE plot
pca.fit %>%
  # extract eigenvalues
```

```
  tidy(matrix = "eigenvalues") %>%
  ggplot(aes(PC, percent)) +
  geom_col() +
  scale_x_continuous(
    # create one axis tick per PC
    breaks = 1:5,

  ) +
  scale_y_continuous(
    name = "Variance Explained",
    # format y axis ticks as percent values
    label = scales::label_percent(accuracy = 1),
    expand = expansion(mult = c(0, 0.06))
  )+
  theme_bw()
```




```
pca.fit %>%
  # add PCs to the original dataset
  augment(volcano_update) %>%
  ggplot(aes(.fittedPC1, .fittedPC2)) +
  xlab("PC1")+
  ylab("PC2")+
  geom_point(aes(color = country ))+
  theme_bw()
```



Discussion: From the PVE plot, we see that the first two principal components explain about 85% variation in the model. Then, from the rotation plot, we can see there are total of 7 arrows and they have roughly equal length. That means the first two PC's are eligible to represent all the features. `population_within_100km`, `population_within_30km`, `population_within_10km`, and `population_within_5km` are mainly described by PC1 and they all pointing to the negative direction of PC1. Notice that both `population_within_10km` and `population_within_5km` are overlapped. That simply means the population within 10 km of active volcano in both China and UK is same as the population within 5 km of that. On the other hand, `longitude`, `latitude`, and `elevation` are mainly explained by PC2 and they all pointing to the positive direction of PC2. That makes sense because `longitude` and `latitude` are geospatial features so these two are highly correlated variables. `elevation` is also another geographical variable and may depend on longitude or latitude.

From the third plot, we can observe that the scatter points that represent China's volcano tend to have high latitude and longitude because those points are being pointed directly by `longitude` and `latitude` arrows in rotation plot. UK on the other hand tend to have none of these features since all the scatter points those represent UK's volcano are shown in the negative direction of the arrows. In other words, volcano in UK tends to have small value in longitude and latitude. We can confirm that from our `volcano_update` dataset (we can access that by `View(volcano_update)`) where we can see UK's volcano have negative value in latitude and longitude whereas China's volcano have positive values.