第一次实验作业

15 MARCH 2021

> 实验背景
>
> > 使用`R`语言内置的`iris`数据集进行分析，掌握基本的数据帧相关操作、简单的统计与绘图等，为日后实验打下基础。
>
> 实验方法
>
> > 使用的`iris`数据集包含了鸢尾属三种植物（`setosa`，`versicolor`，`virginica`）各150个样本的花瓣、萼片的长宽测量值。
> >
> > 首先获取关于数据集的基本信息（大小、列名及其他简单统计）；
> >
> > 其次探索性数据分析，挑选特定关系的数据或通过探索发现有分析价值的数据；
> >
> > 最后选择特定数据关系，作深入分析，得到有价值的结论或推断。
>
> 运行环境（见下）

```
platform          x86_64-pc-linux-gnu
arch              x86_64
os                linux-gnu
system            x86_64, linux-gnu
R version         4.0.4 (2021-02-15)
nickname          Lost Library Book
radian version    0.5.10
python version    3.7.3
```

比较不同品种鸢尾属植物花瓣和萼片的差异

数据集透视

全部数据

```
iris <- tibble(iris)
DT::datatable(iris)
```

列名

`colnames`函数返回所有列名

```
cat(colnames(iris))
```

```
Sepal.Length Sepal.Width Petal.Length Petal.Width Species
```

数据集大小

`dim`函数返回一个

1×2 的关于数据帧的行数与列数

```
cat(dim(iris))
```

```
150 5
```

各观测值简单统计

`summary`函数对各列数据作基本统计，并对数值类变量返回常见的统计量，对字符类返回统计值

```
iris %>%
    summary()
```

```
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

符号	说明
`Min.`	最小值
`1st Qu.`	第一四分位（前 25%）
`Median`	中位数
`Mean`	均值
`3rd Qu.`	第三四分位（后 25%）
`Max.`	最大值
`setosa`等	各字符串类型变量个数统计

样本中萼片最大长度

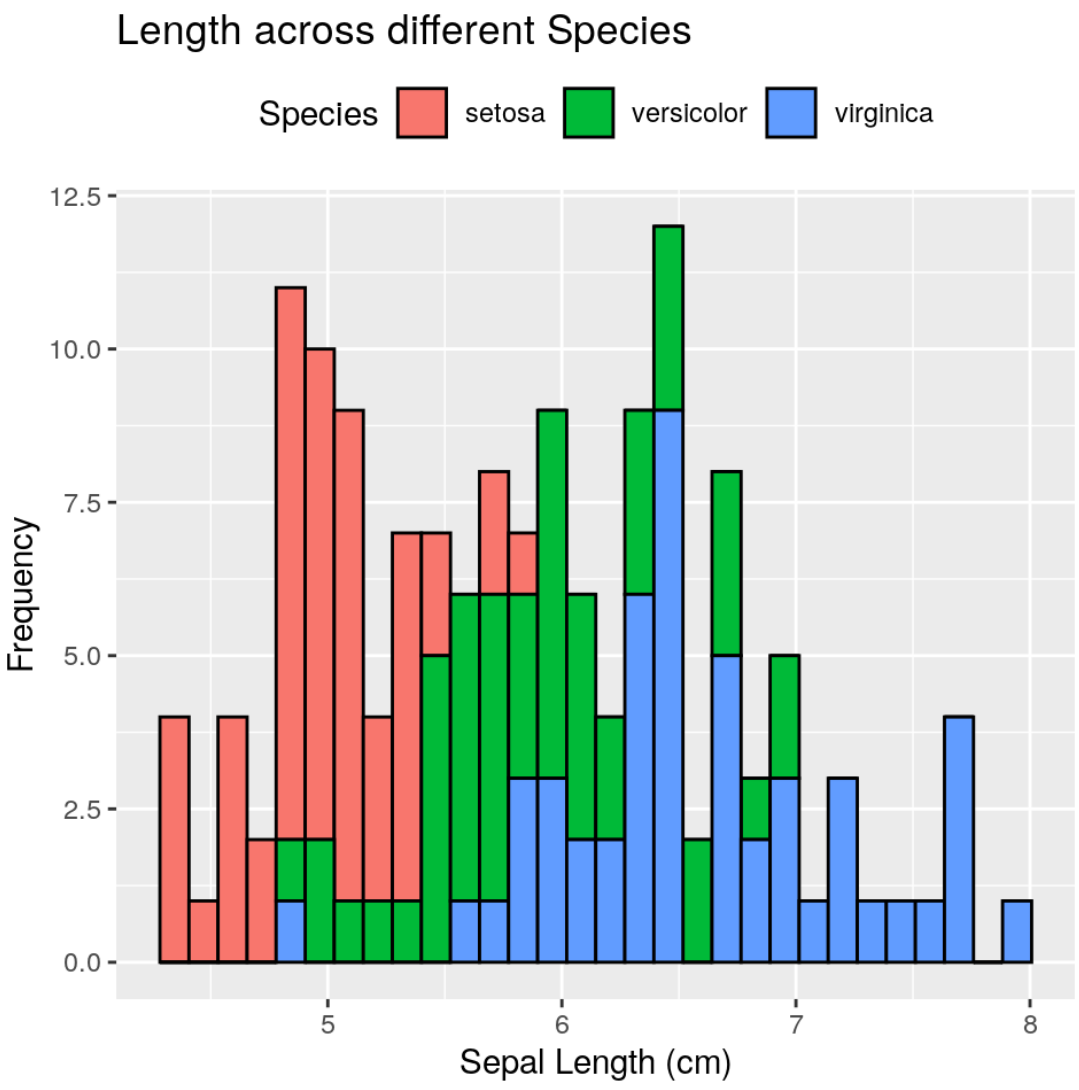
使用`arrange`对`tibble`类型基于特定列排序，使用`head`获取最前6行结果

```
iris %>%
    arrange(desc(Sepal.Length)) %>%
    head() %>%
    DT::datatable()
```

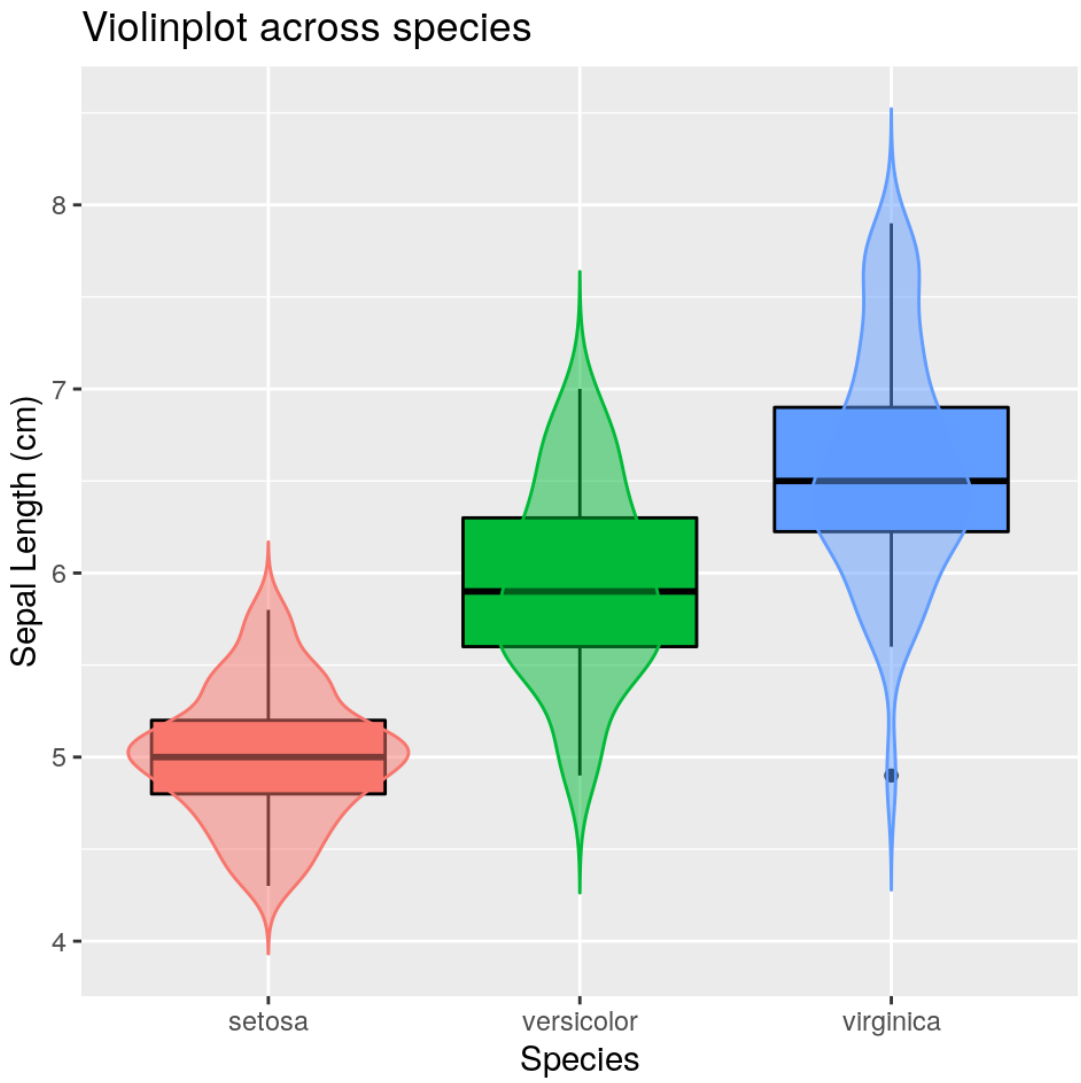
简单统计图

使用`ggplot2`包中函数，绘制各萼片长度的频次统计图与各种类的萼片长度分布提亲图/箱线图

```
library(ggplot2)
iris %>%
    ggplot(aes(x = Sepal.Length, group = Species, fill = Species)) +
    geom_histogram(color = "black") +
    theme(
        legend.position = "top"
    ) +
    labs(
        x = "Sepal Length (cm)",
        y = "Frequency",
        title = "Length across different Species"
    )
```



```r
iris %>%
    ggplot(aes(x = Species, y = Sepal.Length, fill = Species)) +
    geom_boxplot(color = "black") +
    geom_violin(
        trim = FALSE,
        alpha = 0.5,
        scale = "count",
        aes(color = Species)
    ) +
    theme(
        legend.position = "none"
    ) +
    labs(
        x = "Species",
        y = "Sepal Length (cm)",
        title = "Violinplot across species"
    )
```



箱线图中包含了均值（箱图内粗黑线）、四分位线（箱图上下侧黑线），与其他相关量（异常值点、第一至第三四分位范围的外拓1.5倍）；

而提琴图则将数据点分布的密度反应于对应取值处图像宽度，能比箱线图提供更细致的分布信息。

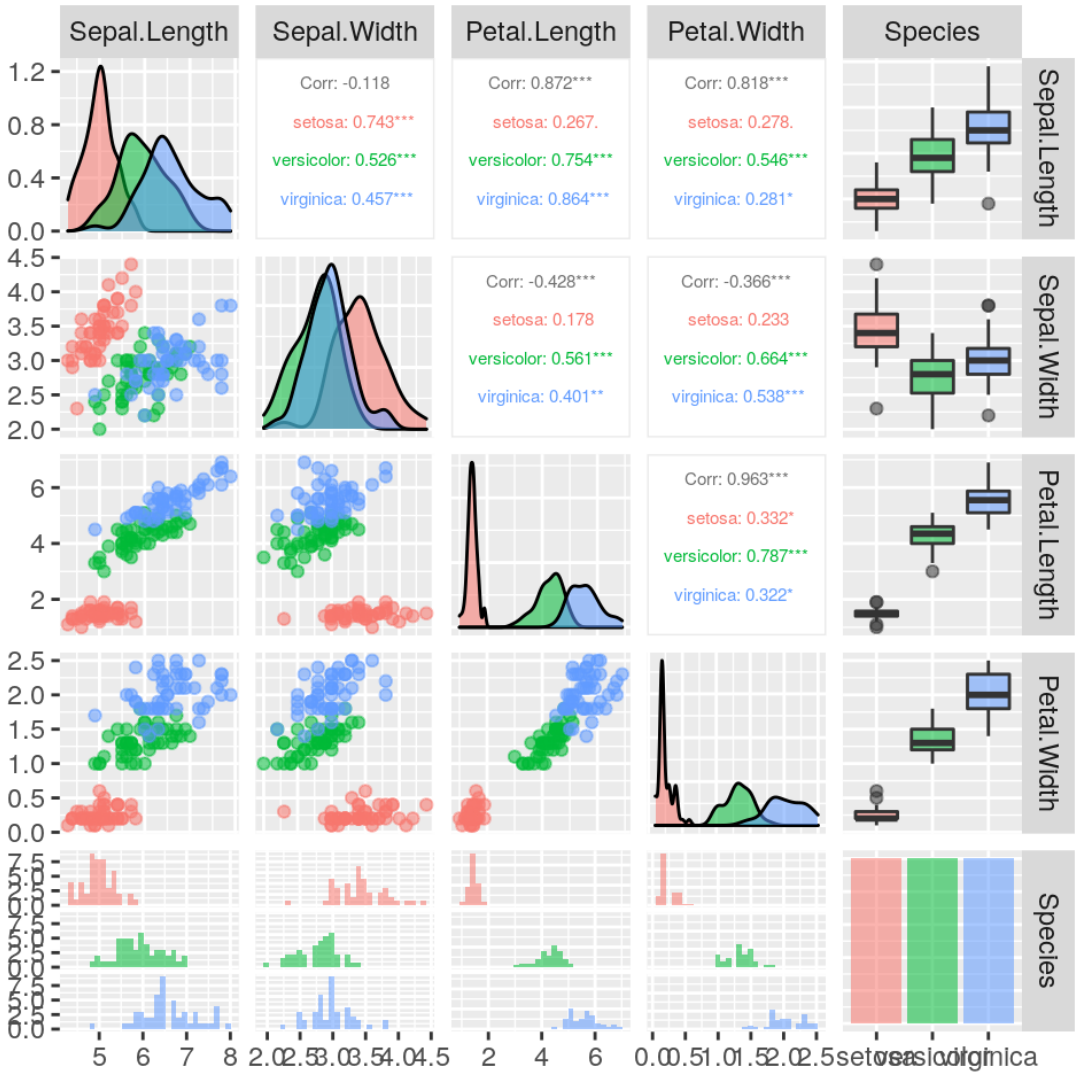
数据相关性探索

数据相关性探索首先通过定向或不定向的大范围粗略相关性分析来发现潜在的数据相关性，基于此发现在后续进行深入分析，得到具有意义的结论

探索性数据分析（EDA）

`GGally`包中包含了`ggpairs`函数，比内置的`pairs`函数提供了更精美的作图、对非数值型统计量更好的兼容处理与更丰富的信息

```
library(GGally)
ggpairs(
    iris,
    aes(col = Species, alpha = 0.1),
    upper = list(continuous = wrap("cor", size = 2))
)
```



不难发现，图中展现的“花瓣长-花瓣宽”、“花瓣长-萼片长”、“花瓣宽-萼片长”等具有非常显著的线性关系（$p < 0.001$），而如“萼片长-萼片宽”则线性关系不明显（$p > 0.1$）。

花瓣与萼片的长度关系

选择性分析花瓣长度与萼片长度之间的线性关系。`R`语言内置了`lm`(linear model)方法能够对给定数据进行线性拟合，后续用`ggplot2`包中函数绘制数据点与拟合结果（黑线）。

```r
lm_fit <- lm(data = iris, formula = Petal.Length ~ Sepal.Length)
summary.lm(lm_fit)
```

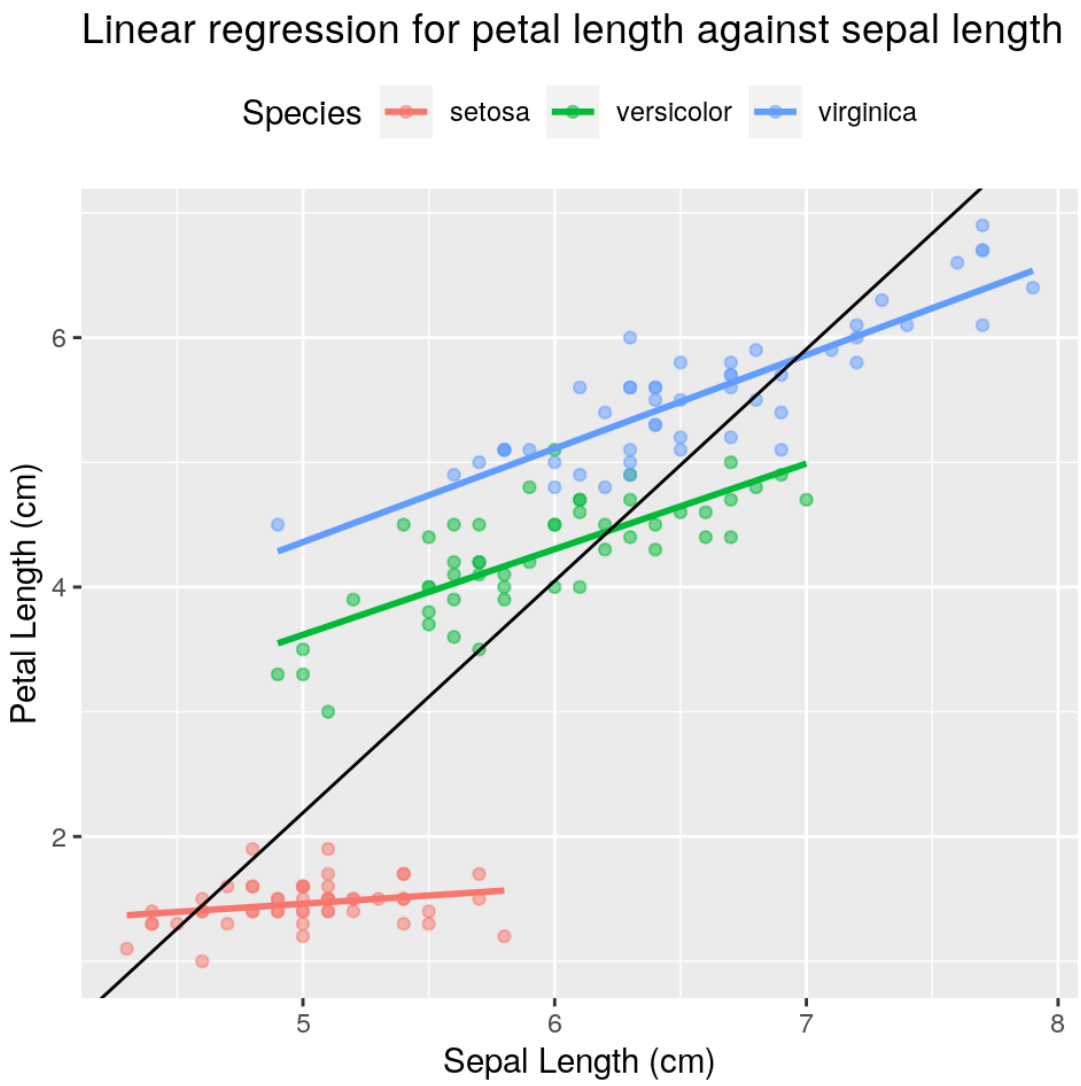
```
Call:
lm(formula = Petal.Length ~ Sepal.Length, data = iris)

Residuals:
     Min       1Q   Median       3Q      Max
-2.47747 -0.59072 -0.00668  0.60484  2.49512

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -7.10144    0.50666  -14.02   <2e-16 ***
Sepal.Length  1.85843    0.08586   21.65   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8678 on 148 degrees of freedom
Multiple R-squared:   0.76, Adjusted R-squared:  0.7583
F-statistic: 468.6 on 1 and 148 DF,  p-value: < 2.2e-16
```

```r
iris %>%
    ggplot(aes(
        x = Sepal.Length,
        y = Petal.Length,
        color = Species
    )) +
    geom_point(
        alpha = 0.5
    ) +
    geom_smooth(
        method = "lm",
        fill = NA,
        alpha = 0.8
    ) +
    geom_abline(
        slope = coef(lm_fit)[[2]],
        intercept = coef(lm_fit)[[1]]
    ) +
    theme(
        legend.position = "top"
    ) +
    labs(
        x = "Sepal Length (cm)",
        y = "Petal Length (cm)",
        title = "Linear regression for petal length against sepal length"
    )
```



不难发现，尽管花瓣长度对萼片长度的数据能够较好地拟合到一条直线上，但在组内进行线性拟合的结果仍会与不分组的结果有较大出入，提示了萼片长与花瓣长的关系在三个组之间显著差异，可以作为聚类分组的参考。

不同种之间萼片长度差异显著性分析 ANOVA

`ANOVA`（方差分析，此处特指`F-test`）和`t-test`最大区别在于前者零假设为所有组间均值相等，而后者的零假设为观测均值为常数（单变量）或两变量均值相同（双变量）。由于此处鸢尾属下样本种类数大于二，故选用`ANOVA`进行假设检验。

```
iris %>%
    aov(Sepal.Length ~ Species, .) %>%
    summary()
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
Species       2  63.21  31.606   119.3 <2e-16 ***
Residuals   147  38.96   0.265
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

结果中

$p < 0.001$，提示组内的均值有非常显著的差异。

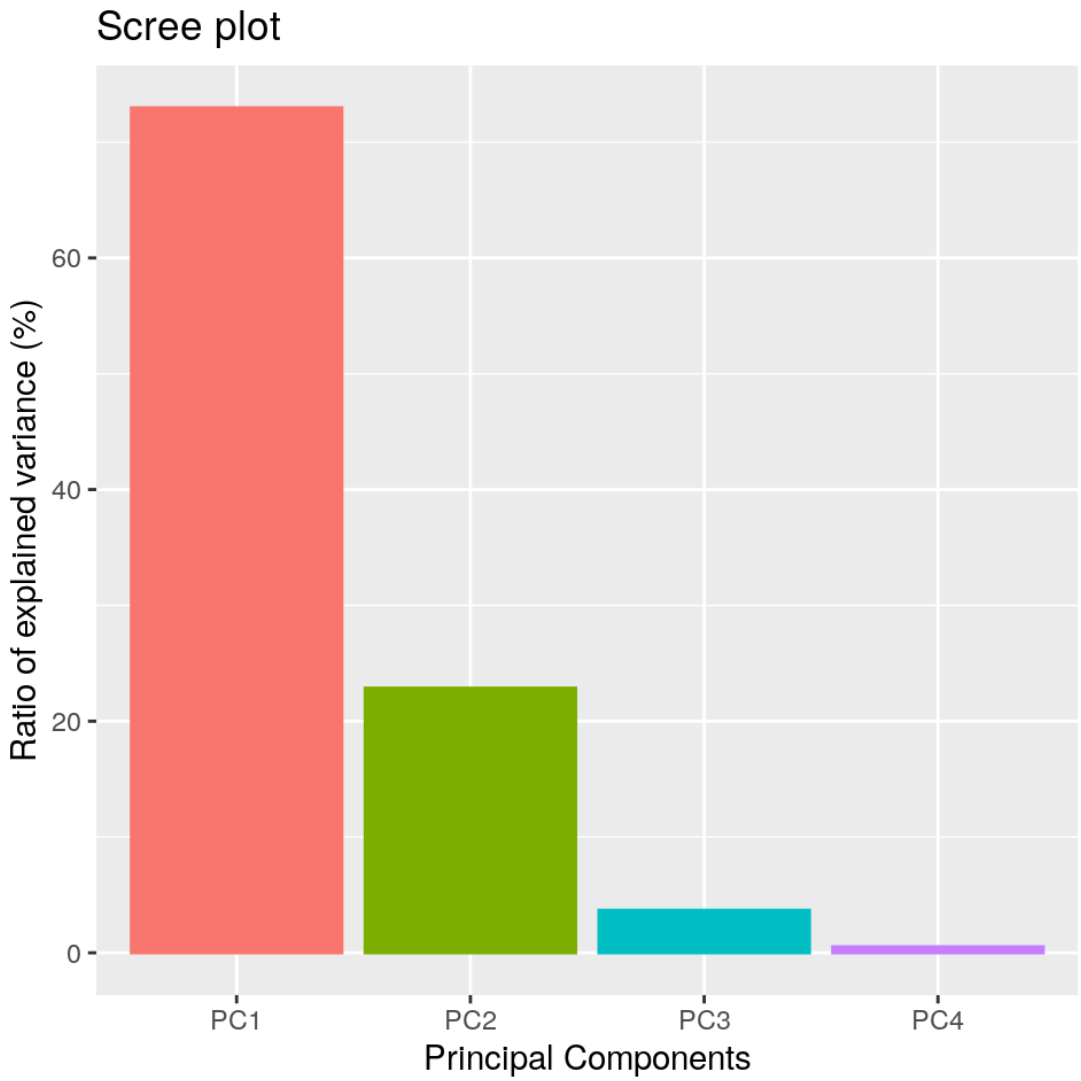
基于四项观测值主成分分析（PCA）进行分类

基于上述分析，可以得知不同的鸢尾属植物在此四项观测内有显著的组间差异，因此可以尝试使用PCA（主成分分析法）提取出特征向量，基于四项观测值对数据进行分类。

PCA碎石图

碎石图主要展现了各主成分（PC）能够解释的原数据点之间的方差比例。

```
library(ggalt)
library(ggfortify)
pca_res <- prcomp(iris[c(1:4)], scale = TRUE, center = TRUE)
pcavar <- pca_res$sdev^2 %>% as_tibble()
pcavar <- pcavar / sum(pcavar) * 100
pcs <- colnames(pca_res$rotation)
pcavp <- data.frame(pcs = pcs, value = pcavar)
pcavp %>%
    ggplot(aes(
        x = pcs,
        y = value,
        fill = pcs,
        color = pcs
    )) +
    geom_col() +
    theme(
        legend.position = "none"
    ) +
    labs(
        x = "Principal Components",
        y = "Ratio of explained variance (%)",
        title = "Scree plot"
    )
```

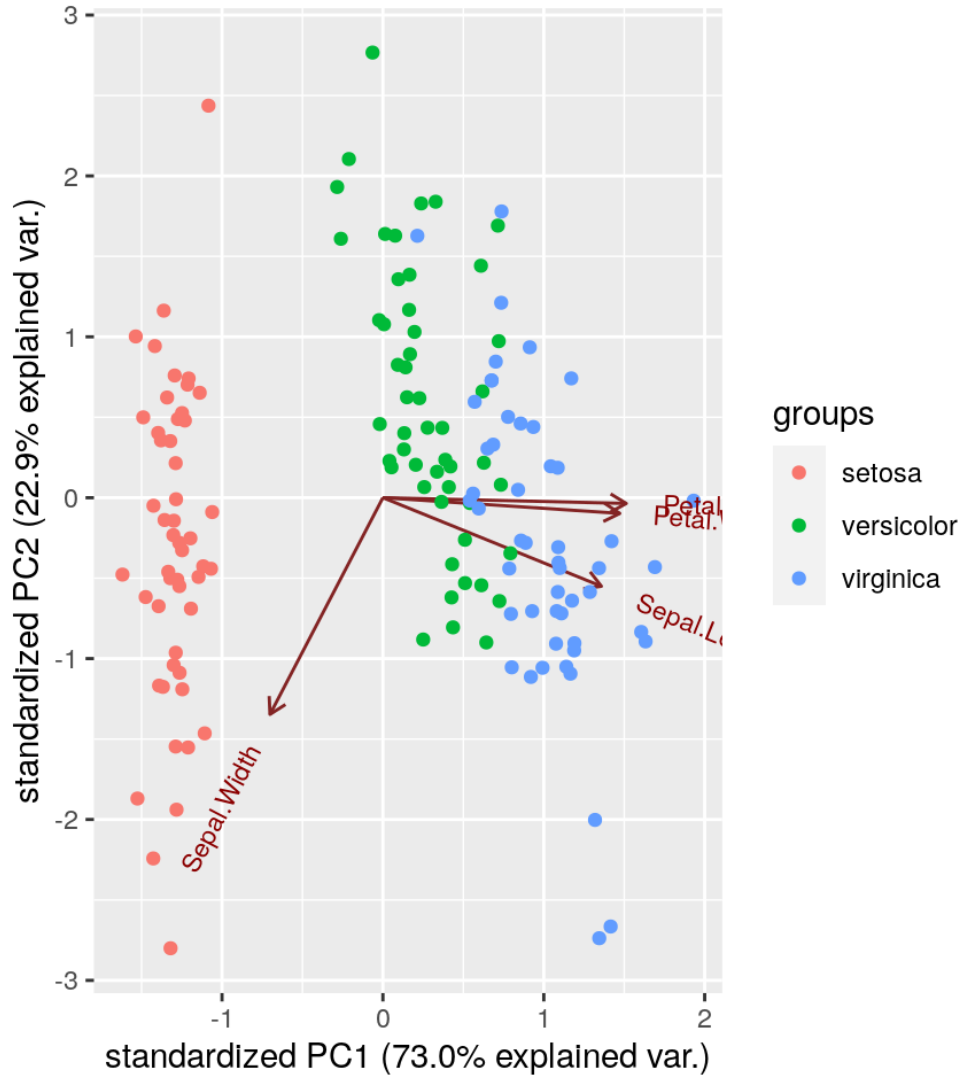


由上图可以看出，前两项主成分所解释的部分占总的95%以上，因此我们只需要选用前两项主成分便能较好的描述数据集的大部分特征，也就能将三类较好地区分开。

主成分作图

`ggfortify` 包中 `ggbiplot` 函数能够比较直观地绘制出 `PCA` 分析的结果，我们可以通过其中的函数来直观观察主成分与各观测值及数据点之间的关系。

```
ggbiplot(
    pca_res,
    groups = iris$Species
)
```

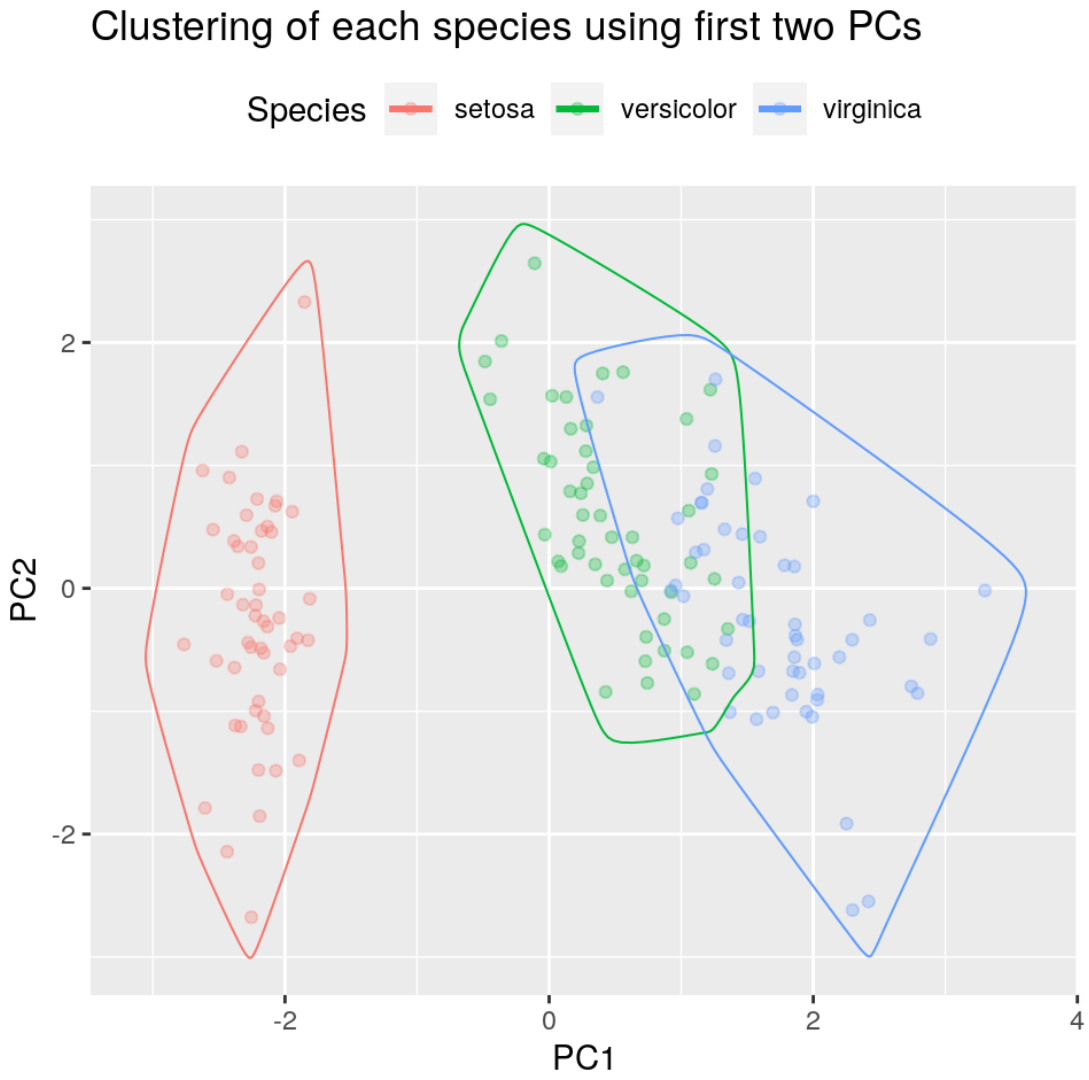


我们还可以输出各主成分与各观测值之间的相关性：

```
DT::datatable(pca_res$rotation)
```

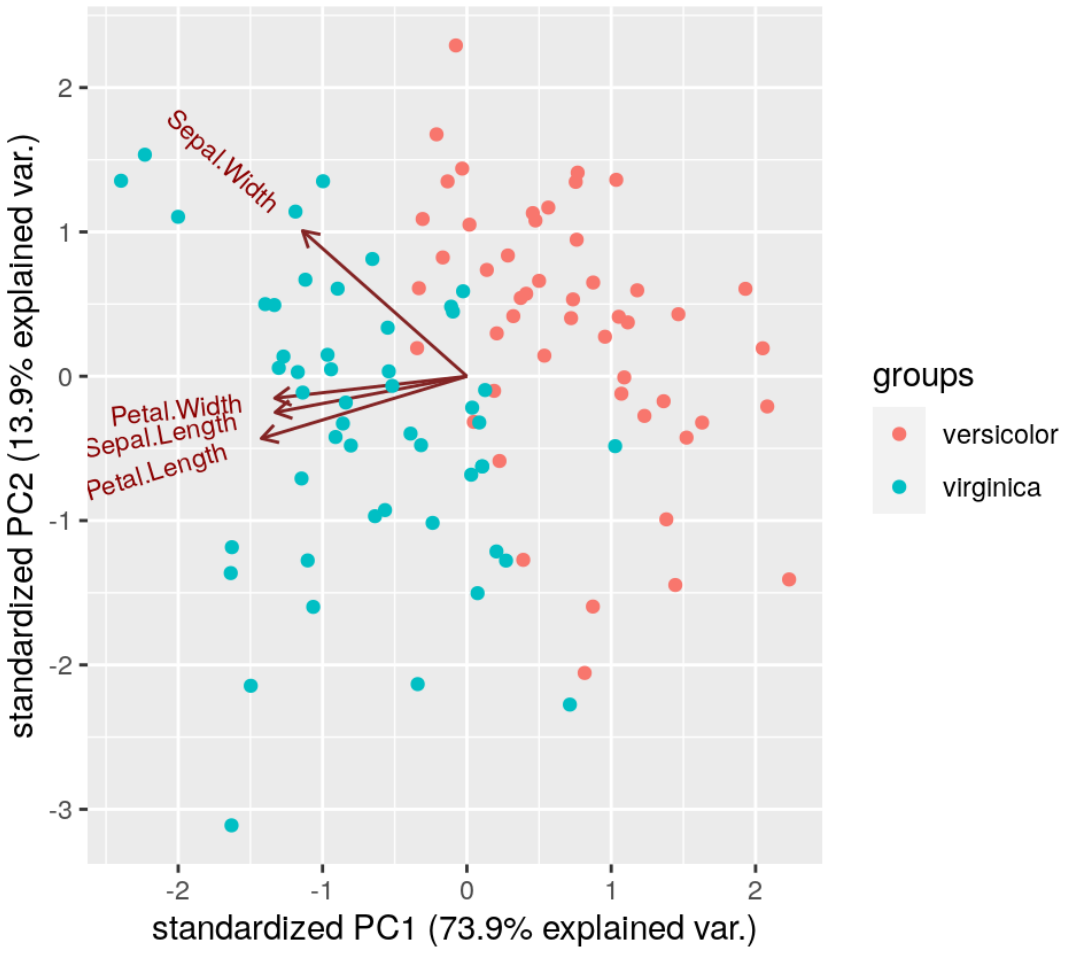
由于数据点之间经过主成分分解后有一定重合，故我们调用`ggalt`包中`geom_encircle`函数绘制各分类的包络曲线（`ggbiplot`中`ellipse`参数有类似效果，其可绘制椭圆，但由于其描述的数据范围并不非常贴合，故不用于此处），以此曲线来观察分类重合情况。

```r
pca_plot_data <- data.frame(pca_res$x, Species = iris$Species)
pca_plot_data %>%
    ggplot(aes(
        x = PC1,
        y = PC2,
        color = Species
    )) +
    geom_point(
        alpha = 0.3
    ) +
    geom_encircle(
        data = pca_plot_data[pca_plot_data$Species == "setosa", ],
        aes(x = PC1, y = PC2),
        s_shape = 0.8
    ) +
    geom_encircle(
        data = pca_plot_data[pca_plot_data$Species == "versicolor", ],
        aes(x = PC1, y = PC2),
        s_shape = 0.8
    ) +
    geom_encircle(
        data = pca_plot_data[pca_plot_data$Species == "virginica", ],
        aes(x = PC1, y = PC2),
        s_shape = 0.8
    ) +
    scale_y_continuous(expand = c(0.1, 0.1)) +
    scale_x_continuous(expand = c(0.1, 0.1)) +
    labs(
        x = "PC1",
        y = "PC2",
        title = "Clustering of each species using first two PCs"
    ) +
    theme(
        legend.position = "top"
    )
```

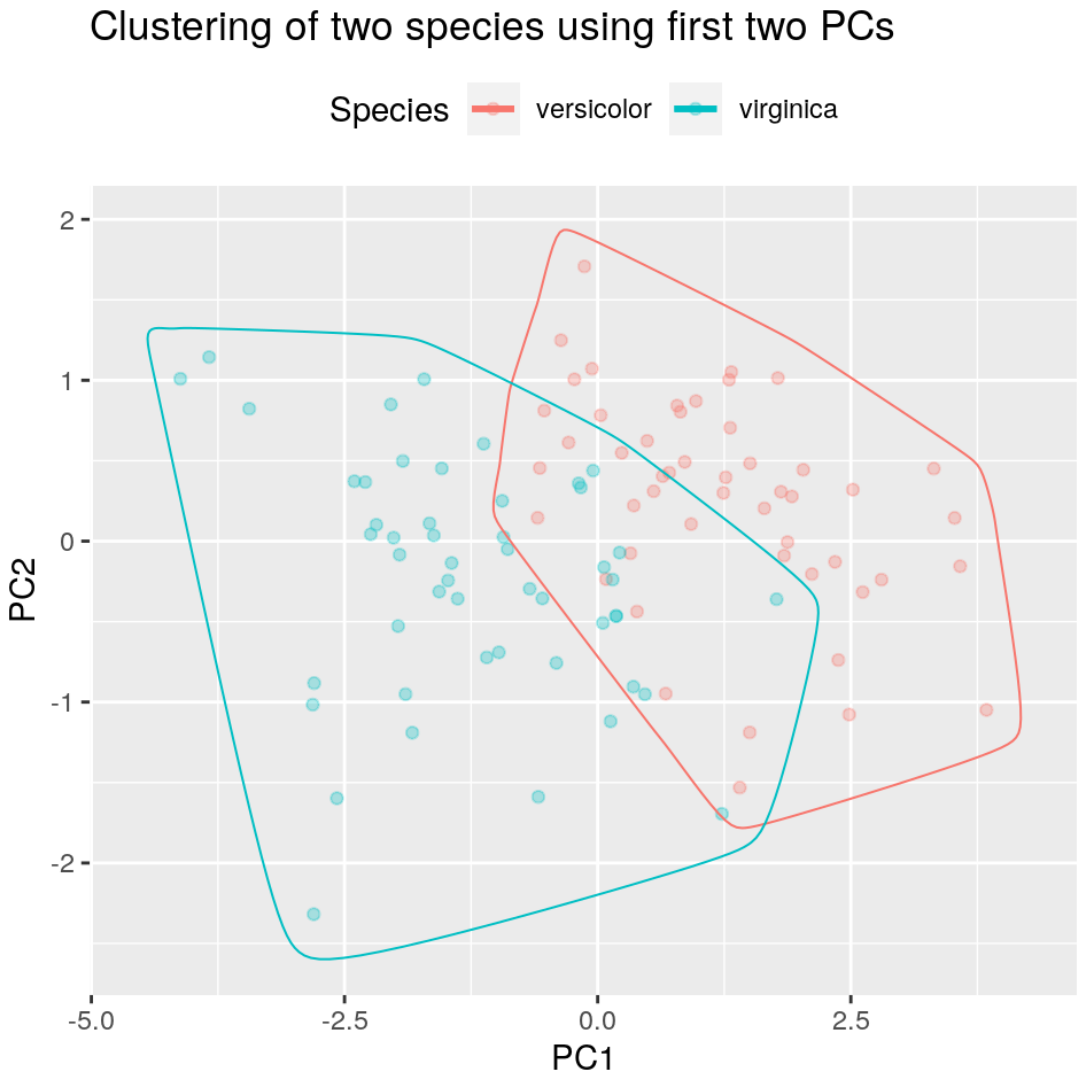


由上图可见，结果中`setosa`组被非常好的分离出来，而其余的`versicolor`和`virginica`组仍有较大范围的重合，这既有数据集样本点本身就比较接近的原因，也反过来说明了单纯使用`PCA`方法来对数据进行分类不一定有非常好的效果，可以考虑使用其他的方法进行分类，或者尝试将此二组单独进行`PCA`分类：

```
filtered <- iris[iris$Species == "versicolor" | iris$Species == "virginica", ]
pca_res_extra <- prcomp(
    filtered[c(1:4)],
    scale = TRUE,
    center = TRUE
)
ggbiplot(
    pca_res_extra,
    groups = filtered$Species
)
```




```r
pca_plot_data_ext <- data.frame(pca_res_extra$x, Species = filtered$Species)
pca_plot_data_ext %>%
    ggplot(aes(
        x = PC1,
        y = PC2,
        color = Species
    )) +
    geom_point(
        alpha = 0.3
    ) +
    geom_encircle(
        data = pca_plot_data_ext[pca_plot_data_ext$Species == "versicolor", ],
        aes(x = PC1, y = PC2),
        s_shape = 0.8
    ) +
    geom_encircle(
        data = pca_plot_data_ext[pca_plot_data_ext$Species == "virginica", ],
        aes(x = PC1, y = PC2),
        s_shape = 0.8
    ) +
    scale_y_continuous(expand = c(0.1, 0.1)) +
    scale_x_continuous(expand = c(0.1, 0.1)) +
    labs(
        x = "PC1",
        y = "PC2",
        title = "Clustering of two species using first two PCs"
    ) +
    theme(
        legend.position = "top"
    )
```



此处得到的结果中重合部分的数据点个数已经明显少于之前的结果，说明针对重合较多的部分进行二次`PCA`是有效果的；但仍有相当部分重合，说明该数据集中此二种植物的数据并不能单纯通过`PCA`进行完美的区分，还需要使用或结合其他的方法，也进一步说明了`PCA`的局限性。

参考资料

1. https://www.iuj.ac.jp/faculty/kucc625/method/anova.html
2. https://www.datacamp.com/community/tutorials/pca-analysis-r
3. https://www.rdocumentation.org/packages/ggfortify/versions/0.4.11