R_hw7

Sihyuan Han

Baltimore City Crime Data

- 1-1,2

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
bc_crime <-
  read_csv(file = "../R_data/BPD_Part_1_Victim_Based_Crime_Data.csv")
```

```
## Parsed with column specification:
## cols(
##   CrimeDate = col_character(),
##   CrimeTime = col_time(format = ""),
##   CrimeCode = col_character(),
##   Location = col_character(),
##   Description = col_character(),
##   `Inside/Outside` = col_character(),
##   Weapon = col_character(),
##   Post = col_double(),
##   District = col_character(),
##   Neighborhood = col_character(),
##   Longitude = col_double(),
##   Latitude = col_double(),
##   `Location 1` = col_logical(),
##   Premise = col_character(),
##   vri_name1 = col_character(),
##   `Total Incidents` = col_double()
## )
```

```
# str(bc_crime)
nrow(bc_crime)
```

```
## [1] 316623
```

```
tail(bc_crime)
```

```
## # A tibble: 6 x 16
##   CrimeDate CrimeTime CrimeCode Location Description ‘Inside/Outside‘ Weapon
##   <chr>     <time>    <chr>     <chr>    <chr>       <chr>            <chr>
## 1 01/01/19~ 10:30     2A        1900 AR~ RAPE        I                OTHER
## 2 05/01/19~ 00:01     2A        600 W 3~ RAPE        I                OTHER
## 3 06/01/19~ 00:00     2A        4400 OL~ RAPE        I                OTHER
## 4 07/01/19~ 23:00     2A        4000 SP~ RAPE        I                OTHER
## 5 07/20/19~ 21:00     2A        5400 RO~ RAPE        <NA>             OTHER
## 6 10/30/19~ 00:00     2A        3100 FE~ RAPE        I                OTHER
## # ... with 9 more variables: Post <dbl>, District <chr>, Neighborhood <chr>,
## #   Longitude <dbl>, Latitude <dbl>, ‘Location 1‘ <lgl>, Premise <chr>,
## #   vri_name1 <chr>, ‘Total Incidents‘ <dbl>
```

- 1-3

```
bc_crime%>%
  mutate(CrimeDate = parse_date(CrimeDate, format = "%m/%d/%Y"),
         CrimeCode = parse_factor(CrimeCode),
         Description = parse_factor(Description),
         ‘Inside/Outside‘ = parse_factor(‘Inside/Outside‘),
         Weapon = parse_factor(Weapon),
         District = parse_factor(District)) ->
  bc_crime
head(bc_crime)
```

```
## # A tibble: 6 x 16
##   CrimeDate  CrimeTime CrimeCode Location Description ‘Inside/Outside‘ Weapon
##   <date>     <time>    <fct>     <chr>    <fct>       <fct>            <fct>
## 1 2020-10-03 04:35     4E        2700 GA~ COMMON ASS~ I                <NA>
## 2 2020-10-03 02:30     5D        400 RUS~ BURGLARY    O                <NA>
## 3 2020-10-03 13:27     4B        4300 BE~ AGG. ASSAU~ O                KNIFE
## 4 2020-10-03 08:17     6C        6600 BE~ LARCENY     I                <NA>
## 5 2020-10-03 00:50     4B        700 E 2~ AGG. ASSAU~ I                KNIFE
## 6 2020-10-03 04:06     5D        1700 MA~ BURGLARY    I                <NA>
## # ... with 9 more variables: Post <dbl>, District <fct>, Neighborhood <chr>,
## #   Longitude <dbl>, Latitude <dbl>, ‘Location 1‘ <lgl>, Premise <chr>,
## #   vri_name1 <chr>, ‘Total Incidents‘ <dbl>
```

- 1-4

```
bc_crime%>%
  rename(Inside_Outside = ‘Inside/Outside‘,
         Location_1 = ‘Location 1‘,
         Total_Incidents = ‘Total Incidents‘) ->
  bc_crime

bc_crime%>%
  select(Inside_Outside, Location_1, Total_Incidents)
```

```
## # A tibble: 316,623 x 3
##    Inside_Outside Location_1 Total_Incidents
##    <fct>          <lgl>                <dbl>
##  1 I              NA                       1
##  2 O              NA                       1
##  3 O              NA                       1
##  4 I              NA                       1
##  5 I              NA                       1
##  6 I              NA                       1
##  7 I              NA                       1
##  8 O              NA                       1
##  9 O              NA                       1
## 10 O              NA                       1
## # ... with 316,613 more rows
```

- 1-5

```
# Check duplicated rows
bc_crime%>%
  summarize(dist = nrow(distinct(.)))
```

```
## # A tibble: 1 x 1
##     dist
##    <int>
## 1 303953
```

```
nrow(bc_crime)
```

```
## [1] 316623
```

```
# How many duplicated rows?
316623-303953
```

```
## [1] 12670
```

```
# Remove duplicated rows
bc_crime%>%
  distinct(.keep_all = TRUE) ->
  bc_crime
```

- 1-6

```
bc_crime%>%
  summarize(across(everything(), ~sum(is.na(.))))
```

```
## # A tibble: 1 x 16
##   CrimeDate CrimeTime CrimeCode Location Description Inside_Outside Weapon  Post
##       <int>     <int>     <int>    <int>       <int>          <int>  <int> <int>
## 1         0        26         0     1548           0              0      0   706
## # ... with 8 more variables: District <int>, Neighborhood <int>,
## #   Longitude <int>, Latitude <int>, Location_1 <int>, Premise <int>,
## #   vri_name1 <int>, Total_Incidents <int>
```

```r
# Which columns have the most and least number of values other than NA?
# max: CrimeDate,CrimeCode,Description,Weapon,District,Total_Incidents
# min: Location_1

# remove column with all NA values
bc_crime%>%
  select_if(~any(!is.na(.)))
```

```
## # A tibble: 303,953 x 15
##    CrimeDate  CrimeTime CrimeCode Location Description Inside_Outside Weapon
##    <date>     <time>    <fct>     <chr>    <fct>       <fct>          <fct>
##  1 2020-10-03 04:35     4E        2700 GA~ COMMON ASS~ I              <NA>
##  2 2020-10-03 02:30     5D        400 RUS~ BURGLARY    O              <NA>
##  3 2020-10-03 13:27     4B        4300 BE~ AGG. ASSAU~ O              KNIFE
##  4 2020-10-03 08:17     6C        6600 BE~ LARCENY     I              <NA>
##  5 2020-10-03 00:50     4B        700 E 2~ AGG. ASSAU~ I              KNIFE
##  6 2020-10-03 04:06     5D        1700 MA~ BURGLARY    I              <NA>
##  7 2020-10-03 05:46     5D        3500 DO~ BURGLARY    I              <NA>
##  8 2020-10-03 06:15     3AJF      2400 BL~ ROBBERY - ~ O              FIREA~
##  9 2020-10-03 02:55     3AJF      1700 BO~ ROBBERY - ~ O              FIREA~
## 10 2020-10-03 00:05     3AJF      3500 CL~ ROBBERY - ~ O              FIREA~
## # ... with 303,943 more rows, and 8 more variables: Post <dbl>, District <fct>,
## #   Neighborhood <chr>, Longitude <dbl>, Latitude <dbl>, Premise <chr>,
## #   vri_name1 <chr>, Total_Incidents <dbl>
```

```r
# Extra Credit
bc_crime%>%
  summarize(across(everything(), ~sum(!is.na(.))))
```

```
## # A tibble: 1 x 16
##   CrimeDate CrimeTime CrimeCode Location Description Inside_Outside Weapon
##       <int>     <int>     <int>    <int>       <int>          <int>  <int>
## 1    303953    303927    303953   302405      303953         303953 303953
## # ... with 9 more variables: Post <int>, District <int>, Neighborhood <int>,
## #   Longitude <int>, Latitude <int>, Location_1 <int>, Premise <int>,
## #   vri_name1 <int>, Total_Incidents <int>
```

- 1-7

```r
# head(sort(unique(bc_crime$Inside_Outside)))

bc_crime%>%
  mutate(
    Inside_Outside = case_when(
      Inside_Outside == "I"  ~"Inside",
      Inside_Outside == "O"  ~"Outside",
      TRUE ~  as.character(Inside_Outside)
    )
  ) ->
  bc_crime
head(bc_crime)
```

```
## # A tibble: 6 x 16
##   CrimeDate  CrimeTime CrimeCode Location Description Inside_Outside Weapon
##   <date>     <time>    <fct>     <chr>    <fct>       <chr>          <fct>
## 1 2020-10-03 04:35     4E        2700 GA~ COMMON ASS~ Inside         <NA>
## 2 2020-10-03 02:30     5D        400 RUS~ BURGLARY    Outside        <NA>
## 3 2020-10-03 13:27     4B        4300 BE~ AGG. ASSAU~ Outside        KNIFE
## 4 2020-10-03 08:17     6C        6600 BE~ LARCENY     Inside         <NA>
## 5 2020-10-03 00:50     4B        700 E 2~ AGG. ASSAU~ Inside         KNIFE
## 6 2020-10-03 04:06     5D        1700 MA~ BURGLARY    Inside         <NA>
## # ... with 9 more variables: Post <dbl>, District <fct>, Neighborhood <chr>,
## #   Longitude <dbl>, Latitude <dbl>, Location_1 <lgl>, Premise <chr>,
## #   vri_name1 <chr>, Total_Incidents <dbl>
```

- 1-8

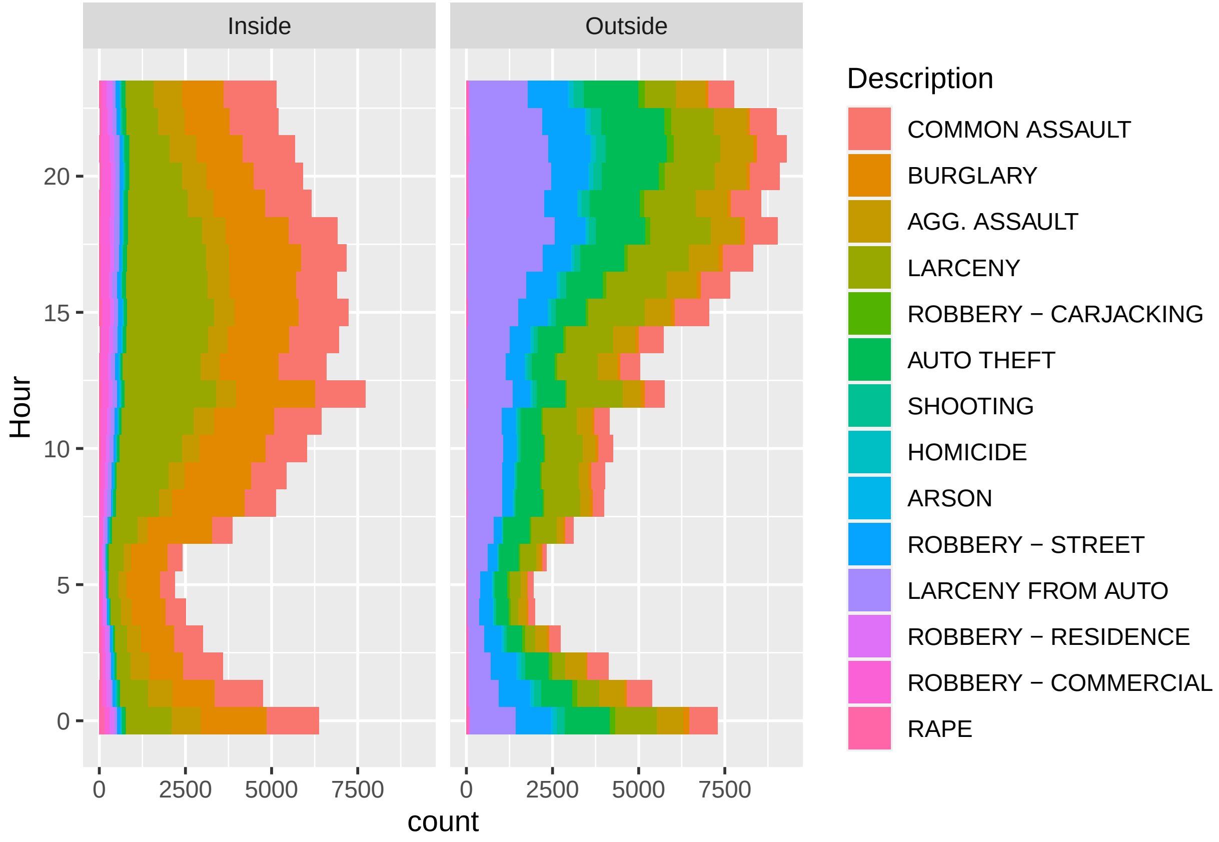
```
bc_crime%>%
  separate(CrimeTime,
           into = c("Hour", "Minute", "Second"),
           sep = ":",
           remove = FALSE,
           convert = TRUE) ->
  bc_crime
head(bc_crime)
```

```
## # A tibble: 6 x 19
##   CrimeDate  CrimeTime  Hour Minute Second CrimeCode Location Description
##   <date>     <time>    <int>  <int>  <int> <fct>     <chr>    <fct>
## 1 2020-10-03 04:35         4     35      0 4E        2700 GA~ COMMON ASS~
## 2 2020-10-03 02:30         2     30      0 5D        400 RUS~ BURGLARY
## 3 2020-10-03 13:27        13     27      0 4B        4300 BE~ AGG. ASSAU~
## 4 2020-10-03 08:17         8     17      0 6C        6600 BE~ LARCENY
## 5 2020-10-03 00:50         0     50      0 4B        700 E 2~ AGG. ASSAU~
## 6 2020-10-03 04:06         4      6      0 5D        1700 MA~ BURGLARY
## # ... with 11 more variables: Inside_Outside <chr>, Weapon <fct>, Post <dbl>,
## #   District <fct>, Neighborhood <chr>, Longitude <dbl>, Latitude <dbl>,
## #   Location_1 <lgl>, Premise <chr>, vri_name1 <chr>, Total_Incidents <dbl>
```

- 1-9

```
bc_crime%>%
  filter(!is.na(Inside_Outside))%>%
  ggplot(aes(y = Hour, fill = Description))+
  geom_bar(width = 1)+
  facet_wrap( ~Inside_Outside)
```

```
## Warning: Removed 26 rows containing non-finite values (stat_count).
```



- 1-10

```
bc_crime%>%
  filter(CrimeTime >= parse_time("00:00:00", format = "%H:%M:%S") &
         CrimeTime <= parse_time("04:00:00", format = "%H:%M:%S")) ->
  bc_crime_qten

round(prop.table(table(bc_crime_qten$Total_Incidents, bc_crime_qten$Inside_Outside, useNA = "ifany"), ma
```

```
##
##     Inside Outside <NA>
##   1   0.42    0.46 0.13
```

- 1-11

```
bc_crime%>%
  saveRDS("../R_output/bc_crime_hw7_output.rds")
```

- Describe the difference in the file sizes: .Rds file is larger than the original file.
- Reload the file you just saved into a variable called balt2 and count the number of rows

```
balt2 <- readRDS("../R_output/bc_crime_hw7_output.rds")
nrow(balt2)
```

```
## [1] 303953
```

Billboard Data

- 2-1

```
data("billboard")
head(billboard)
```

```
## # A tibble: 6 x 79
##   artist track date.entered   wk1   wk2   wk3   wk4   wk5   wk6   wk7   wk8
##   <chr>  <chr> <date>       <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 2 Pac  Baby~ 2000-02-26      87    82    72    77    87    94    99    NA
## 2 2Ge+h~ The ~ 2000-09-02      91    87    92    NA    NA    NA    NA    NA
## 3 3 Doo~ Kryp~ 2000-04-08      81    70    68    67    66    57    54    53
## 4 3 Doo~ Loser 2000-10-21      76    76    72    69    67    65    55    59
## 5 504 B~ Wobb~ 2000-04-15      57    34    25    17    17    31    36    49
## 6 98^0   Give~ 2000-08-19      51    39    34    26    26    19     2     2
## # ... with 68 more variables: wk9 <dbl>, wk10 <dbl>, wk11 <dbl>, wk12 <dbl>,
## #   wk13 <dbl>, wk14 <dbl>, wk15 <dbl>, wk16 <dbl>, wk17 <dbl>, wk18 <dbl>,
## #   wk19 <dbl>, wk20 <dbl>, wk21 <dbl>, wk22 <dbl>, wk23 <dbl>, wk24 <dbl>,
## #   wk25 <dbl>, wk26 <dbl>, wk27 <dbl>, wk28 <dbl>, wk29 <dbl>, wk30 <dbl>,
## #   wk31 <dbl>, wk32 <dbl>, wk33 <dbl>, wk34 <dbl>, wk35 <dbl>, wk36 <dbl>,
## #   wk37 <dbl>, wk38 <dbl>, wk39 <dbl>, wk40 <dbl>, wk41 <dbl>, wk42 <dbl>,
## #   wk43 <dbl>, wk44 <dbl>, wk45 <dbl>, wk46 <dbl>, wk47 <dbl>, wk48 <dbl>,
## #   wk49 <dbl>, wk50 <dbl>, wk51 <dbl>, wk52 <dbl>, wk53 <dbl>, wk54 <dbl>,
## #   wk55 <dbl>, wk56 <dbl>, wk57 <dbl>, wk58 <dbl>, wk59 <dbl>, wk60 <dbl>,
## #   wk61 <dbl>, wk62 <dbl>, wk63 <dbl>, wk64 <dbl>, wk65 <dbl>, wk66 <lgl>,
## #   wk67 <lgl>, wk68 <lgl>, wk69 <lgl>, wk70 <lgl>, wk71 <lgl>, wk72 <lgl>,
## #   wk73 <lgl>, wk74 <lgl>, wk75 <lgl>, wk76 <lgl>
```

- 2-2,3,4

```
billboard%>%
  pivot_longer(cols = c(wk1:wk76),
               names_to = "week",
               values_to = "ranking",
               names_prefix = "wk",
               names_transform = list(week = as.numeric),
               values_drop_na = TRUE)%>%
  mutate(date = date.entered+(week-1)*7)%>%
  separate(date.entered,
           into = c("year", "month", "day"),
           sep = "-")%>%
  select(-month, -day)%>%
  arrange(artist, track, week)%>%
  relocate(year)
```

```
## # A tibble: 5,307 x 6
##    year  artist  track                    week ranking date      
##    <chr> <chr>   <chr>                   <dbl>   <dbl> <date>    
##  1 2000  2 Pac   Baby Don’t Cry (Keep...     1      87 2000-02-26
##  2 2000  2 Pac   Baby Don’t Cry (Keep...     2      82 2000-03-04
##  3 2000  2 Pac   Baby Don’t Cry (Keep...     3      72 2000-03-11
##  4 2000  2 Pac   Baby Don’t Cry (Keep...     4      77 2000-03-18
##  5 2000  2 Pac   Baby Don’t Cry (Keep...     5      87 2000-03-25
##  6 2000  2 Pac   Baby Don’t Cry (Keep...     6      94 2000-04-01
##  7 2000  2 Pac   Baby Don’t Cry (Keep...     7      99 2000-04-08
##  8 2000  2Ge+her The Hardest Part Of ...     1      91 2000-09-02
##  9 2000  2Ge+her The Hardest Part Of ...     2      87 2000-09-09
## 10 2000  2Ge+her The Hardest Part Of ...     3      92 2000-09-16
## # ... with 5,297 more rows
```

Iris dataset

- 3-1,2

```
# read_lines("../R_data/iris.names")
iris_data <- read_csv("../R_data/iris.data",
                      col_names = c(
  "sepal_length", "sepal_width", "petal_length", "petal_width", "species"
  )
                      )
```

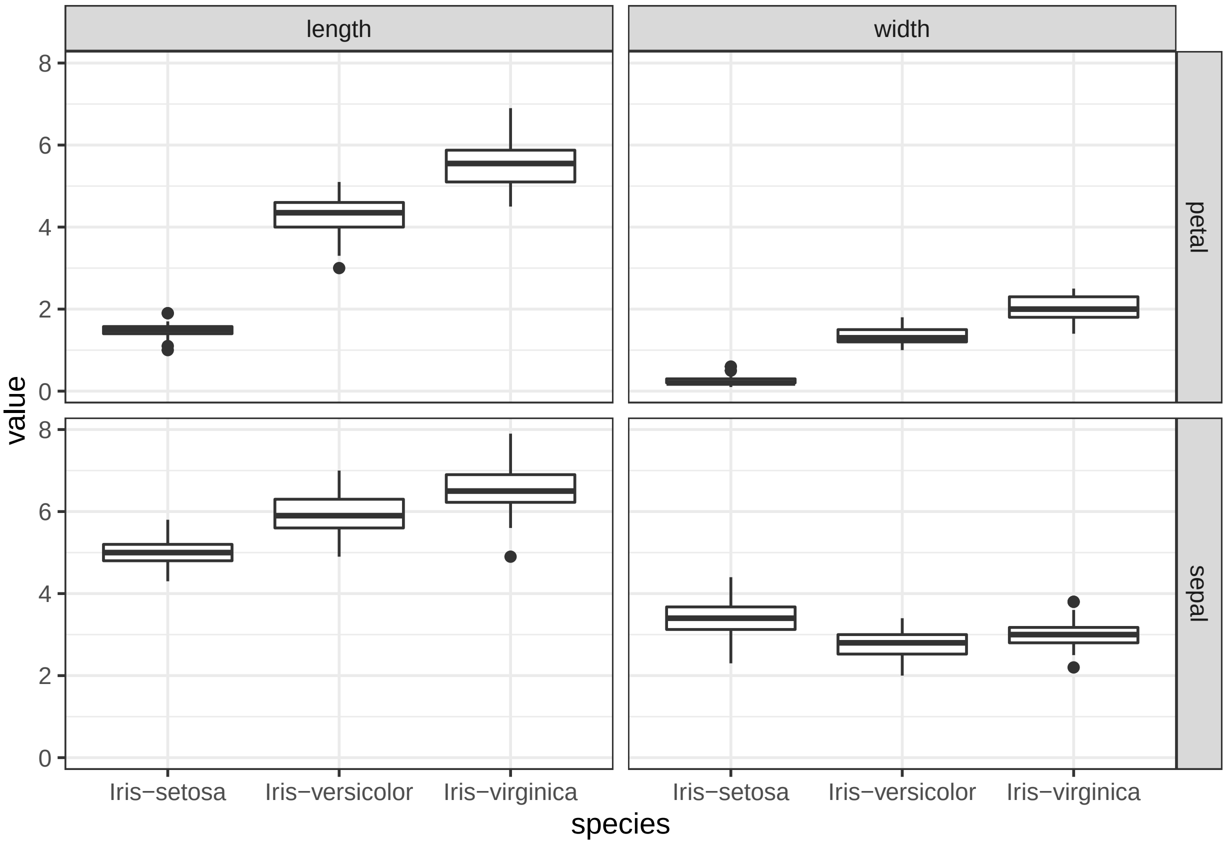
```
## Parsed with column specification:
## cols(
##   sepal_length = col_double(),
##   sepal_width = col_double(),
##   petal_length = col_double(),
##   petal_width = col_double(),
##   species = col_character()
## )
```

```
head(iris_data)
```

```
## # A tibble: 6 x 5
##   sepal_length sepal_width petal_length petal_width species    
##          <dbl>       <dbl>        <dbl>       <dbl> <chr>      
## 1          5.1         3.5          1.4         0.2 Iris-setosa
## 2          4.9         3            1.4         0.2 Iris-setosa
## 3          4.7         3.2          1.3         0.2 Iris-setosa
## 4          4.6         3.1          1.5         0.2 Iris-setosa
## 5          5           3.6          1.4         0.2 Iris-setosa
## 6          5.4         3.9          1.7         0.4 Iris-setosa
```

- 3-3

```r
iris_data%>%
  pivot_longer(cols = c(sepal_length:petal_width),
               names_to = "name",
               values_to = "value")%>%
  separate(name,
           into = c("plant_part","measure_dim"),
           sep = "_")%>%
  ggplot(aes(species, value))+
  geom_boxplot()+
  facet_grid(plant_part~measure_dim)+
  theme_bw()
```


length
width
petal
sepal
value
Iris−setosa
Iris−versicolor
Iris−virginica
Iris−setosa
Iris−versicolor
Iris−virginica
species