Code

Introduction to programming in R and Python by Natalia Kazakova and Andrey Yakovenko

Project 1

Hide

```
install.packages("stringi", dependencies=TRUE, INSTALL_opts = c('--no-lock'))
```

Hide

```
install.packages("stringr", dependencies=TRUE, INSTALL_opts = c('--no-lock'))
```

Hide

```
install.packages("tidyverse")
```

Hide

```
library(tidyverse)
```

1-2-3. Find a dataset, download it and load the data into R

We will study the species repartition and weight of animals caught in plots in our study area. The dataset is stored as a comma separated value (CSV) file. Each row holds information for a single animal, and the columns represent:

record_id - Unique id for the observation month - month of observation day - day of observation year - year of observation plot_id - ID of a particular plot species_id - 2-letter code sex - sex of animal (“M”, “F”) hindfoot_length - length of the hindfoot in mm weight - weight of the animal in grams genus - genus of animal species - species of animal taxa - some taxon, e.g. Rodent, Reptile, Bird, Rabbit plot_type - type of plot

Before the work, we need to check and change (if necessary) our working directory:

Hide

```
getwd()
```

```
[1] "/Users/19820585/Desktop"
```

Hide

```
setwd('/Users/19820585/Documents/hse/Introduction to programming in R and Python')
```

Hide

```
surveys <- read.csv("/Users/19820585/Documents/hse/Introduction to programming in R and Python/portal_data_joined.csv")
```

Let's check the top of this data frame:

	record_id <int>	mo… <int>	… <int>	year <int>	plot_id <int>	species_id <chr>	… <chr>	hindfoot_length <int>	weight <int>
1	1	7	16	1977	2	NL	M	32	*NA*
2	72	8	19	1977	2	NL	M	31	*NA*
3	224	9	13	1977	2	NL		*NA*	*NA*
4	266	10	16	1977	2	NL		*NA*	*NA*
5	349	11	12	1977	2	NL		*NA*	*NA*
6	363	11	12	1977	2	NL		*NA*	*NA*

6 rows | 1-10 of 13 columns

… and size:

Hide

```
dim(surveys)
```

```
[1] 34786    13
```

Let's inspect the structure of a data frame with the function str():

Hide

```
str(surveys)
```

```
'data.frame':   34786 obs. of  13 variables:
 $ record_id      : int  1 72 224 266 349 363 435 506 588 661 ...
 $ month          : int  7 8 9 10 11 11 12 1 2 3 ...
 $ day            : int  16 19 13 16 12 12 10 8 18 11 ...
 $ year           : int  1977 1977 1977 1977 1977 1977 1977 1978 1978 1978 ...
 $ plot_id        : int  2 2 2 2 2 2 2 2 2 2 ...
 $ species_id     : chr  "NL" "NL" "NL" "NL" ...
 $ sex            : chr  "M" "M" "" "" ...
 $ hindfoot_length: int  32 31 NA NA NA NA NA NA NA NA ...
 $ weight         : int  NA NA NA NA NA NA NA NA 218 NA ...
 $ genus          : chr  "Neotoma" "Neotoma" "Neotoma" "Neotoma" ...
 $ species        : chr  "albigula" "albigula" "albigula" "albigula" ...
 $ taxa           : chr  "Rodent" "Rodent" "Rodent" "Rodent" ...
 $ plot_type      : chr  "Control" "Control" "Control" "Control" ...
```

Moreover, it can be useful to check summary statistics for each column:

Hide

```
summary(surveys)
```

```
   record_id         month              day             year         plot_id       s
pecies_id            sex
 Min.   :    1   Min.   : 1.000   Min.   : 1.0   Min.   :1977   Min.   : 1.00   Le
ngth:34786       Length:34786
 1st Qu.: 8964   1st Qu.: 4.000   1st Qu.: 9.0   1st Qu.:1984   1st Qu.: 5.00   Cl
ass :character   Class :character
 Median :17762   Median : 6.000   Median :16.0   Median :1990   Median :11.00   Mo
de  :character   Mode  :character
 Mean   :17804   Mean   : 6.474   Mean   :16.1   Mean   :1990   Mean   :11.34
 3rd Qu.:26655   3rd Qu.:10.000   3rd Qu.:23.0   3rd Qu.:1997   3rd Qu.:17.00
 Max.   :35548   Max.   :12.000   Max.   :31.0   Max.   :2002   Max.   :24.00

 hindfoot_length     weight          genus              species              taxa
plot_type
 Min.   : 2.00   Min.   :  4.00   Length:34786       Length:34786       Length:347
86       Length:34786
 1st Qu.:21.00   1st Qu.: 20.00   Class :character   Class :character   Class :cha
racter   Class :character
 Median :32.00   Median : 37.00   Mode  :character   Mode  :character   Mode  :cha
racter   Mode  :character
 Mean   :29.29   Mean   : 42.67
 3rd Qu.:36.00   3rd Qu.: 48.00
 Max.   :70.00   Max.   :280.00
 NA's   :3348    NA's   :2503
```

4. Pick the variables that you think can be relevant for your analysis. Copy these variables into a separate dataset

We decided to drop two variables from our dataset - taxon and type of plot:

5. Check that all R data types actually match the data (e.g. no numbers are stored as characters). If there are, convert the data to the relevant types

Hide

```
str(surveys_new)
```

```
'data.frame':   34786 obs. of  11 variables:
 $ record_id      : int  1 72 224 266 349 363 435 506 588 661 ...
 $ month          : int  7 8 9 10 11 11 12 1 2 3 ...
 $ day            : int  16 19 13 16 12 12 10 8 18 11 ...
 $ year           : int  1977 1977 1977 1977 1977 1977 1977 1978 1978 1978 ...
 $ plot_id        : int  2 2 2 2 2 2 2 2 2 2 ...
 $ species_id     : chr  "NL" "NL" "NL" "NL" ...
 $ sex            : chr  "M" "M" "" "" ...
 $ hindfoot_length: int  32 31 NA NA NA NA NA NA NA NA ...
 $ weight         : int  NA NA NA NA NA NA NA NA 218 NA ...
 $ genus          : chr  "Neotoma" "Neotoma" "Neotoma" "Neotoma" ...
 $ species        : chr  "albigula" "albigula" "albigula" "albigula" ...
```

It looks fine

6. In this dataset, check if there are any missing data. If there are, clean your data by removing the observations where data are missing

Let’s start by removing observations of animals for which weight and hindfoot_length are missing, or the sex has not been determined:

Because we are interested in plotting how species abundances have changed through time, we are also going to remove observations for rare species (i.e., that have been observed less than 50 times). We will do this in two steps: first, we are going to create a data set that counts how often each species has been observed, and filter out the rare species. Second, we will extract only the observations for these more common species:

Let’s check our new data frame:

Hide

```
dim(surveys_complete)
```

```
[1] 30521    11
```

7. After cleaning the dataset, save your result for further analysis to the RDS format

Also, we will save it to the CSV format

Hide

```
write_csv(surveys_complete, path = "/Users/19820585/Documents/hse/Introduction to programming in R and Python//surveys_complete.csv")
```

Project 3

Hide

```
install.packages("ggfortify")
```

Hide

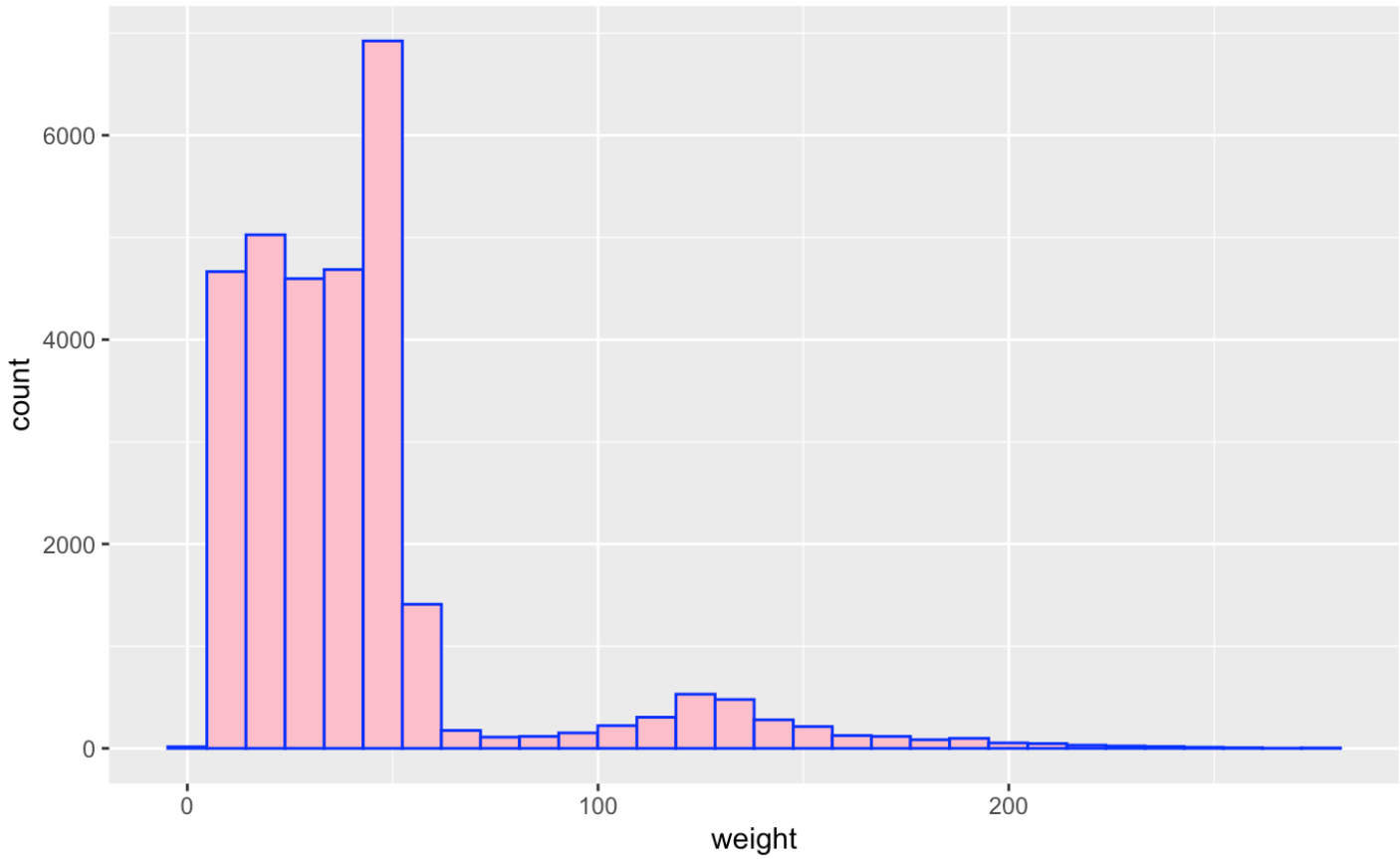
```
library(ggfortify)
```

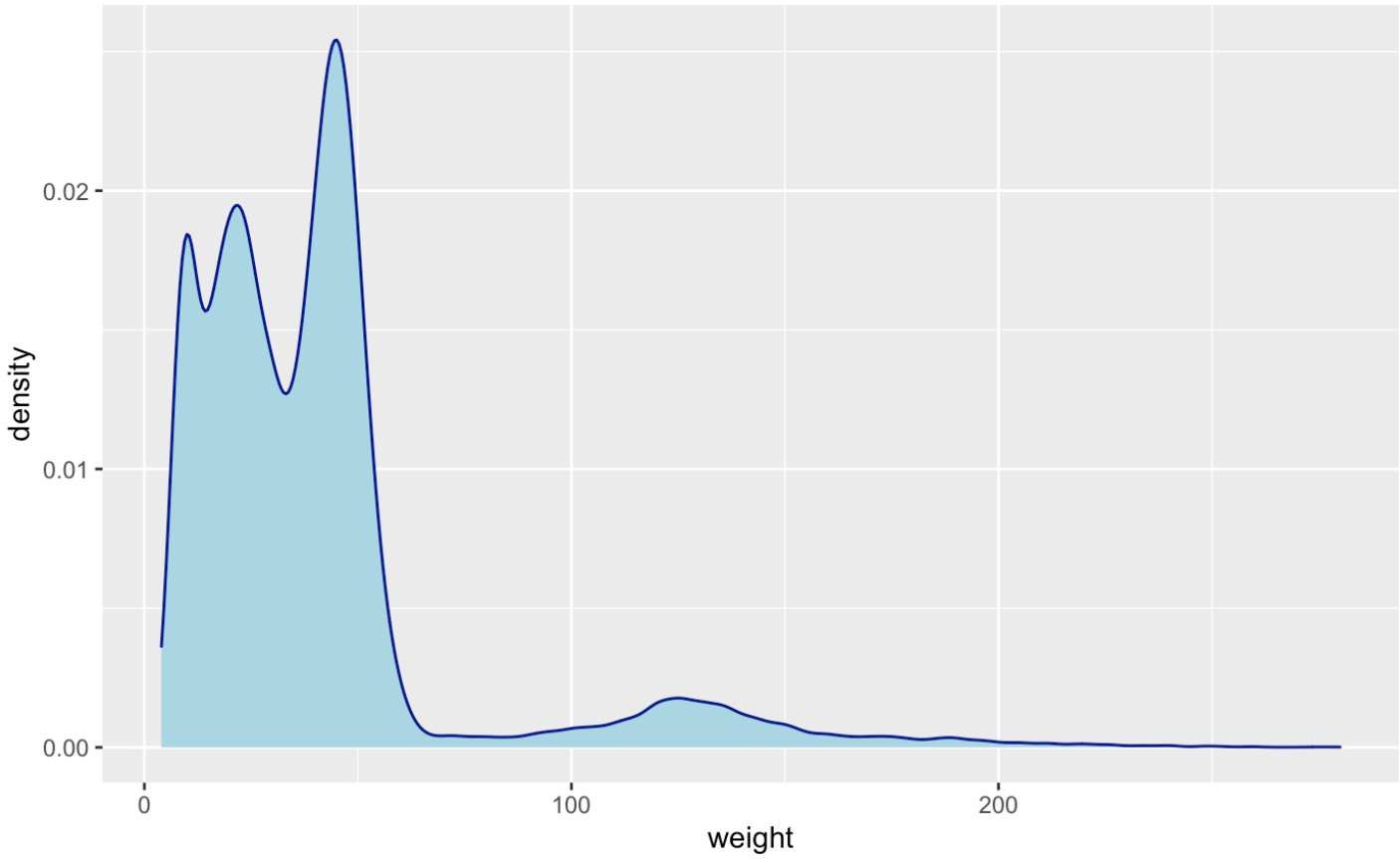
a. A histogram and a density plot of some numeric (ratio or interval scale 😊) variables. These plots show us the distribution of weight within our data set.

Hide

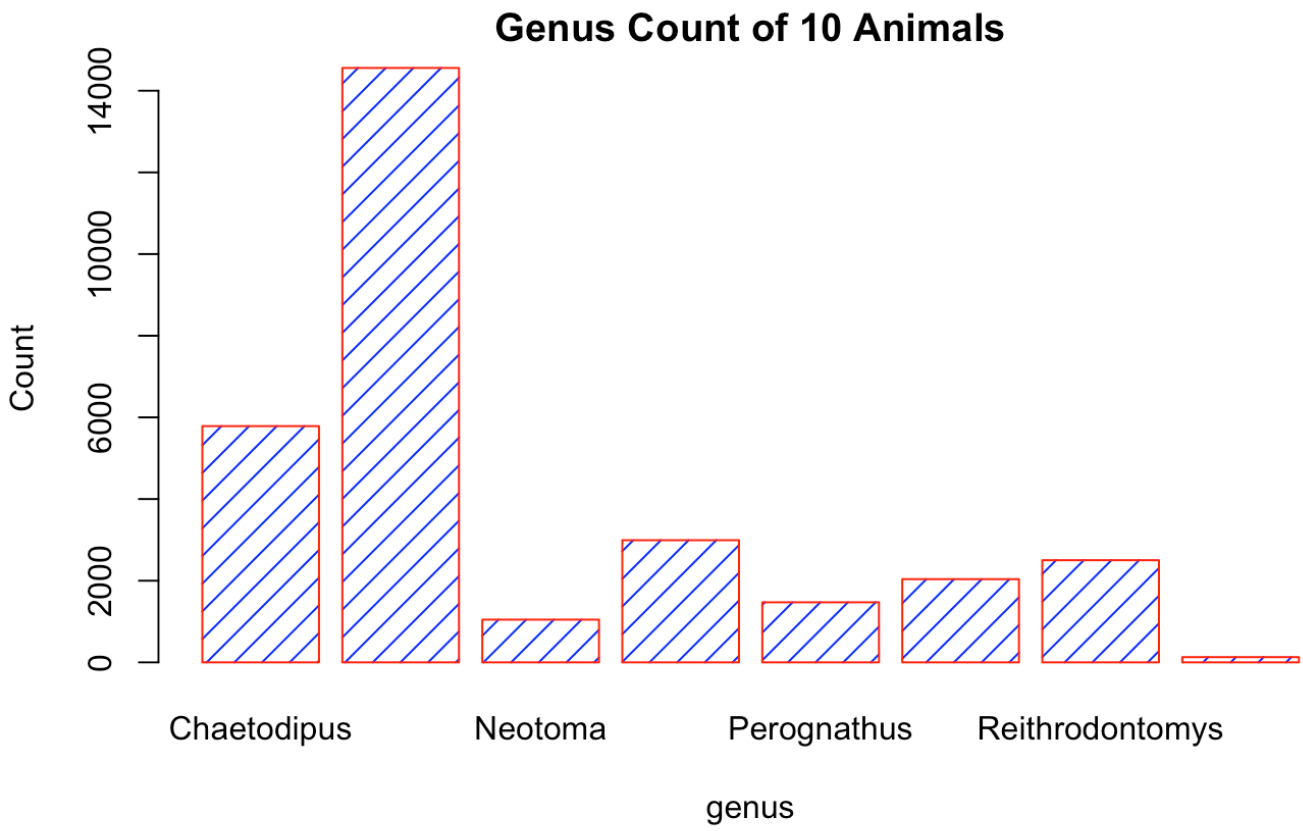
```
ggplot(data=surveys_complete, aes(x=weight))+geom_histogram(color = "blue", fill = "pink")
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```





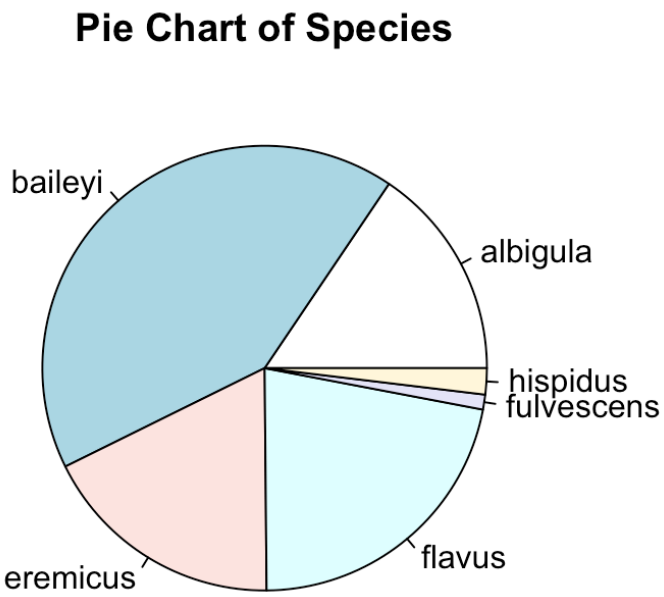
b. A bar and a pie chart for some factor variables (if you do not have factor variables in your dataset, take some numeric variables and convert them to factors)



Hide

```
head(df_pie2)
```

```
df_pie
  albigula    baileyi   eremicus     flavus fulvescens   hispidus
      1046       2808       1200       1471         73        128
```



c. Box (or violin) plots for some numeric variables Here we can see box plots, that reflect animals' weight within their species.