Assignment02

zerofrom

2024-10-05

1. Data Wrangling

1.1 (Q1)

Properties:

```
##  [1] "Month"        "Day"          "Year"         "CaptureTime"  "ReleaseTime"
##  [6] "BandNumber"   "Species"      "Age"          "Sex"          "Wing"
## [11] "Weight"       "Culmen"       "Hallux"       "Tail"         "StandardTail"
## [16] "Tarsus"       "WingPitFat"   "KeelFat"      "Crop"
```

Species names:

```
##   Species
## 1      RT
## 2      CH
## 3      SS
```

Sample Weight data:

```
##   Weight
## 1    920
## 2    930
## 3    990
## 4    470
## 5    170
```

Data Frame hSF:

```
##   Wing Weight Tail
## 1  412   1090  230
## 2  412   1210  210
## 3  405   1120  238
## 4  393   1010  222
## 5  371   1010  217
```

1.1 (Q2)

How many variables does the data frame hSF have?

```
## [1] 3
```

What would you say to communicate this information to a Machine Learning practitioner?

```
## [1] "There are 3 variables in the data frame hSF used for training."
```

How many examples does the data frame hSF have? How many observations? How many cases?

```
## [1] 398
```

1.2 (Q1)

Sort by Wing:

```
##    Wing Weight Tail
## 1  37.2   1180  210
## 2 111.0   1340  226
## 3 199.0   1290  222
## 4 241.0   1320  235
## 5 262.0   1020  200
```

1.3 (Q1)

hawkSpeciesNameCodes:

```
##   species_code species_name_full
## 1           CH          Cooper's
## 2           RT        Red-tailed
## 3           SS     Sharp-shinned
```

1.3 (Q2)

hawksFullName:

```
##   Month Day Year CaptureTime ReleaseTime  BandNumber Age Sex Wing Weight Culmen
## 1     9  19 1992       13:30               877-76317   I      385    920   25.7
## 2     9  22 1992       10:30               877-76318   I      376    930     NA
## 3     9  23 1992       12:45               877-76319   I      381    990   26.7
## 4     9  23 1992       10:50               745-49508   I   F  265    470   18.7
## 5     9  27 1992       11:15              1253-98801   I   F  205    170   12.5
## 6     9  28 1992       11:25             1207-55910    I      412   1090   28.5
## 7     9  28 1992       13:30               877-76320   I      370    960   25.3
##   Hallux Tail StandardTail Tarsus WingPitFat KeelFat Crop       Species
## 1   30.1  219           NA     NA         NA      NA   NA    Red-tailed
## 2     NA  221           NA     NA         NA      NA   NA    Red-tailed
## 3   31.3  235           NA     NA         NA      NA   NA    Red-tailed
## 4   23.5  220           NA     NA         NA      NA   NA      Cooper's
## 5   14.3  157           NA     NA         NA      NA   NA Sharp-shinned
## 6   32.2  230           NA     NA         NA      NA   NA    Red-tailed
## 7   30.1  212           NA     NA         NA      NA   NA    Red-tailed
```

1.3 (Q3)

hawksFullName Select Print:

```
##         Species Wing Weight
## 1    Red-tailed  385    920
## 2    Red-tailed  376    930
## 3    Red-tailed  381    990
## 4      Cooper's  265    470
## 5 Sharp-shinned  205    170
## 6    Red-tailed  412   1090
```

```
## 7    Red-tailed  370    960
```

Does it matter what type of join function you use here? In what situations would it make a difference?

```
## [1] "A left_join B: Return all rows in A."
```

```
## [1] "A right_join B: Return all rows in B."
```

```
## [1] "A inner_join B: Return only the rows in both A and B based on the specified keys."
```

```
## [1] "A full_join B: Return all rows from both A and B"
```

1.4 (Q1)

bird BMI:

```
##   Species  bird_BMI
## 1      RT 852.69973
## 2      RT 108.75741
## 3      RT  32.57493
## 4      RT  22.72688
## 5      CH  22.40818
## 6      RT  19.54932
## 7      CH  15.21998
## 8      RT  14.85927
```

....

1.5 (Q1)

Summarize_data:

```
## # A tibble: 3 x 6
##   Species       num_rows mn_wing nd_wing t_mn_wing b_wt_ratio
##   <chr>            <int>   <dbl>   <dbl>     <dbl>      <dbl>
## 1 Cooper's            70    244.     240      243.       1.67
## 2 Red-tailed         577    383.     384      385.       3.16
## 3 Sharp-shinned      261    185.     191      184.       1.67
```

1.5 (Q2)

Summarize_na_number:

```
## # A tibble: 3 x 9
##   Species        Wing Weight Culmen Hallux  Tail StandardTail Tarsus  Crop
##   <chr>         <int>  <int>  <int>  <int> <int>        <int>  <int> <int>
## 1 Cooper's          1      0      0      0     0           19     62    21
## 2 Red-tailed        0      5      4      3     0          250    538   254
## 3 Sharp-shinned     0      5      3      3     0           68    233    68
```

....

2. Random experiments, events and sample spaces, and the set theory

2.1 (Q1)

Random experiments: A random experiment is a procedure that meet both of the following conditions:

(1) has a well-defined set of possible outcomes;

(2) could (at least in principle) be repeated arbitrarily many times.

Events: An event is a set of possible outcomes of an experiment.

Sample spaces: A sample space is the set of possible outcomes of interest for a random experiment.

The set theory: A set is just a collection of objects of interest, such as the possible outcomes.

2.1 (Q2)

Example: The result of both rolls of the dice is the same: {(1,1),(2,2),(3,3),(4,4),(5,5),(6,6)}

Sample space: {

(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),

(2,1),(2,2),(2,3),(2,4),(2,5),(2,6),

. . . . . . (6,1),(6,2),(6,3),(6,4),(6,5),(6,6)}

Total number: 2ˆ36

Is the empty set considered as an event: Yes, it represents the event where no outcome occurs, meaning an impossible event.

2.2 (Q1)

$$A \cup B = \{1, 2, 3, 4, 6\}$$

$$A \cup C = \{1, 2, 3, 4, 5, 6\}$$

$$A \cap B = \{2\}$$

$$A \cap C = \emptyset$$

$$A \setminus B = \{1, 3\}$$

$$A \setminus C = \{1, 2, 3\}$$

A and B are not disjoint
A and C are disjoint
B and A B are disjoint
Partition into two sets: {{1, 3, 5}, {2, 4, 6}}
Partition into three sets: {{1, 6}, {2, 3}, {4, 5}}

2.2 (Q1)

1.

$$(A^c)^c = A$$

2.

$$\Omega^c = \emptyset$$

3. Let $x \in B^c$, then $x \notin B$.
   Since $A \subseteq B$,then $x \notin A$,so $x \in A^c$.
   Therefore,$B^c \subseteq A^c$.

4.

$$(A \cap B)^c = \{x \in \Omega : x \notin A \cap B\}$$

   It means that x is either not in A or not in B.
   Therefore,

$$(A \cap B)^c = A^c \cup B^c$$

   Let's suppose a sequence of events $A_1, A_2, \ldots, A_K \subseteq \Omega$.
   The general form of this law for the intersection of multiple sets is:

$$\left(\bigcap_{k=1}^{K} A_k \right)^c = \bigcup_{k=1}^{K} A_k^c$$

5. The complement of the union $A \cup B$ is the set of all elements not in $A \cup B$:

$$(A \cup B)^c = \{x \in \Omega : x \notin A \cup B\}$$

   This means that x is neither in A nor in B.
   This means that

$$x \in A^c \quad \text{and} \quad x \in B^c$$

   This means that

$$x \in A^c \cap B^c$$

   Therefore

$$(A \cup B)^c = A^c \cap B^c$$

6. the complement of the union of multiple sets is the intersection of their complements:

$$\left(\bigcup_{k=1}^{K} A_k \right)^c = \bigcap_{k=1}^{K} A_k^c$$

   This is a generalization of the law for multiple events.

2.2 (Q3)

For each element $w_i \in \Omega$, when forming a subset A, we have two choices:
1. Include the element w_i.
2. Exclude the element w_i.

This means that for every element, there are two possible outcomes.
Since there are K elements in Ω, the total number of subsets can be calculated as follows:

$$|E| = 2^K$$

2.2 (Q4)

1. the empty set $A = \emptyset$ is a valid choice. The intersection of the empty set with any other set is empty, so it is disjoint from all other sets.
2.

. . . .

3. Probability theory

3 (Q1)

3 (Q2)

. . . ..