Week 2 Assignment - Data Cleaning

AUTHOR
Liane Chen

```
library(tidyverse)
library(lubridate)
library(RColorBrewer)
```

```
datadir_raw <- "data/raw/"
datadir_processed <- "data/processed/"

snowcover_data <- read_csv(file.path(datadir_raw, "ASDN_Snow_survey.csv"))
```

```
Rows: 42830 Columns: 11
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr (10): Site, Date, Plot, Location, Snow_cover, Water_cover, Land_cover, T...
dbl  (1): Year

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
snowcover_data_fixed <- snowcover_data %>%
  mutate(snow_days = ifelse(Snow_cover > 10, 1, 0),
         Date2 = as_date(Date))
```

```
Warning: There was 1 warning in `mutate()`.
ℹ In argument: `Date2 = as_date(Date)`.
Caused by warning:
!  72 failed to parse.
```

```
snowcover_data_fixed <- snowcover_data_fixed %>%
  mutate(Date2 = ifelse(is.na(Date2),  dmy("08/06/06"), Date2))

snowcover_data_fixed
```

```
# A tibble: 42,830 × 13
   Site   Year Date      Plot  Location Snow_cover Water_cover Land_cover
   <chr> <dbl> <chr>     <chr> <chr>    <chr>      <chr>       <chr>
 1 barr   2011 29-May-11 brw1  b10      90         0           10
 2 barr   2011 29-May-11 brw1  b12      100        0           0
 3 barr   2011 29-May-11 brw1  b2       90         0           10
 4 barr   2011 29-May-11 brw1  b4       100        0           0
 5 barr   2011 29-May-11 brw1  b6       95         0           5
 6 barr   2011 29-May-11 brw1  b8       95         0           5
 7 barr   2011 29-May-11 brw1  d10      95         0           5
 8 barr   2011 29-May-11 brw1  d12      90         0           10
 9 barr   2011 29-May-11 brw1  d2       95         0           5
10 barr   2011 29-May-11 brw1  d4       95         0           5
# ℹ 42,820 more rows
# ℹ 5 more variables: Total_cover <chr>, Observer <chr>, Notes <chr>,
#   snow_days <dbl>, Date2 <dbl>
```

```r
snowcover_data_fixed <- snowcover_data %>%
  mutate(Date = ifelse(Date == "8&9 june 06", "8 june 06", Date),
         Date2 = dmy(Date))
```

1. Clean the `Water_cover` column to transform it into the correct data type and respect expectations for a percentage

```r
snowcover_data_fixed %>%
  count(Water_cover) %>%
  filter(is.na(as.numeric(Water_cover)))
```

```
Warning: There was 1 warning in `filter()`.
ℹ In argument: `is.na(as.numeric(Water_cover))`.
Caused by warning:
! NAs introduced by coercion

# A tibble: 5 × 2
  Water_cover     n
  <chr>       <int>
1 -              10
2 .             575
3 n/a            32
4 unk             1
5 <NA>          149
```

```r
snowcover_data_fixed %>%
  filter(Water_cover ==".") %>%
  View()
```

```r
snowcover_data_fixed <- snowcover_data_fixed %>%
  mutate(Water_cover = ifelse(Water_cover==".", NA, Water_cover))
```

```r
snowcover_data_fixed <- snowcover_data_fixed %>%
  mutate(Water_cover = as.numeric(Water_cover))
```

```
Warning: There was 1 warning in `mutate()`.
ℹ In argument: `Water_cover = as.numeric(Water_cover)`.
Caused by warning:
! NAs introduced by coercion
```

```
glimpse(snowcover_data_fixed)
```

```
Rows: 42,830
Columns: 12
$ Site        <chr> "barr", "barr", "barr", "barr", "barr", "barr", "barr", "b…
$ Year        <dbl> 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011…
$ Date        <chr> "29-May-11", "29-May-11", "29-May-11", "29-May-11", "29-Ma…
$ Plot        <chr> "brw1", "brw1", "brw1", "brw1", "brw1", "brw1", "brw1", "b…
$ Location    <chr> "b10", "b12", "b2", "b4", "b6", "b8", "d10", "d12", "d2", …
$ Snow_cover  <chr> "90", "100", "90", "100", "95", "95", "95", "90", "95", "9…
$ Water_cover <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 10, 0, 0, 0, 0, 0, 0, 0, 0, …
$ Land_cover  <chr> "10", "0", "10", "0", "5", "5", "5", "10", "5", "5", "0", …
$ Total_cover <chr> "100", "100", "100", "100", "100", "100", "100", "100", "1…
$ Observer    <chr> "adoll", "adoll", "adoll", "adoll", "adoll", "adoll", "ado…
$ Notes       <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…
$ Date2       <date> 2011-05-29, 2011-05-29, 2011-05-29, 2011-05-29, 2011-05-2…
```

2. Clean the `Land_cover` column to transform it into the correct data type and respect expectations for a percentage If 2% or less of your data is weird, getting rid of it is fine. Anything more, something is off. Just document it. Get rid of the lines with negative percentages. For lines with NA Snow cover and -100 Land Cover, that doesn't look right either.

```
snowcover_data_fixed %>%
  count(Land_cover) %>%
  filter(is.na(as.numeric(Land_cover)))
```

```
Warning: There was 1 warning in `filter()`.
ℹ In argument: `is.na(as.numeric(Land_cover))`.
Caused by warning:
! NAs introduced by coercion

# A tibble: 5 × 2
  Land_cover     n
  <chr>      <int>
1 -             10
2 .            585
3 n/a           32
4 unk            1
5 <NA>         144
```

```
snowcover_data_fixed %>%
  filter(Land_cover ==".") %>%
  View()
```

```
snowcover_data_fixed <- snowcover_data_fixed %>%
  mutate(Land_cover = ifelse(Land_cover==".", NA, Land_cover))
```

```r
#get rid of negative percentages (these may be input errors)
snowcover_data_fixed <- snowcover_data_fixed %>%
  mutate(Land_cover = ifelse(Land_cover=="<1", "0", Land_cover))
```

```r
#get rid of the row with "ukn"
snowcover_data_fixed <- snowcover_data_fixed %>%
  mutate(Land_cover = ifelse(Land_cover=="unk", NA, Land_cover))
```

```r
#fix rows with n/a and convert to NA to match others
snowcover_data_fixed <- snowcover_data_fixed %>%
  mutate(Land_cover = ifelse(Land_cover=="n/a", NA, Land_cover))
```

```r
snowcover_data_fixed <- snowcover_data_fixed %>%
  mutate(Land_cover = as.numeric(Land_cover))
```

```
Warning: There was 1 warning in `mutate()`.
ℹ In argument: `Land_cover = as.numeric(Land_cover)`.
Caused by warning:
! NAs introduced by coercion
```

```r
glimpse(snowcover_data_fixed)
```

```
Rows: 42,830
Columns: 12
$ Site        <chr> "barr", "barr", "barr", "barr", "barr", "barr", "barr", "b…
$ Year        <dbl> 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011…
$ Date        <chr> "29-May-11", "29-May-11", "29-May-11", "29-May-11", "29-Ma…
$ Plot        <chr> "brw1", "brw1", "brw1", "brw1", "brw1", "brw1", "brw1", "b…
$ Location    <chr> "b10", "b12", "b2", "b4", "b6", "b8", "d10", "d12", "d2", …
$ Snow_cover  <chr> "90", "100", "90", "100", "95", "95", "95", "90", "95", "9…
$ Water_cover <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 10, 0, 0, 0, 0, 0, 0, 0, 0, …
$ Land_cover  <dbl> 10, 0, 10, 0, 5, 5, 5, 10, 5, 5, 0, 10, 5, 10, 20, 10, 5, …
$ Total_cover <chr> "100", "100", "100", "100", "100", "100", "100", "100", "1…
$ Observer    <chr> "adoll", "adoll", "adoll", "adoll", "adoll", "adoll", "ado…
$ Notes       <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…
$ Date2       <date> 2011-05-29, 2011-05-29, 2011-05-29, 2011-05-29, 2011-05-2…
```

```r
snowcover_data_fixed %>%
  filter(Land_cover > 100)
```

```
# A tibble: 0 × 12
# ℹ 12 variables: Site <chr>, Year <dbl>, Date <chr>, Plot <chr>,
#   Location <chr>, Snow_cover <chr>, Water_cover <dbl>, Land_cover <dbl>,
#   Total_cover <chr>, Observer <chr>, Notes <chr>, Date2 <date>
```

`{r} # write_csv(snowcover_data_fixed, file.path(datadir_processed, "snow_cover.csv")) #`

3. Use the relationship between the three cover columns (Snow, Water, Land) to infer missing values where possible and recompute the `Total_cover` column