Visualization of Missing Values and the Effect of Different Weather Variables

Anton Hung

2022-11-01

```
setwd('/Volumes/GoogleDrive/Mon disque/wrangling/project/wranglinghub')
source('functions_library/functions_library.R')
```

```
## Loading required package: Matrix
```

```
## Loading required package: carData
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x tidyr::expand() masks Matrix::expand()
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x tidyr::pack()   masks Matrix::pack()
## x dplyr::recode() masks car::recode()
## x purrr::some()   masks car::some()
## x tidyr::unpack() masks Matrix::unpack()
```

Loading the data

```
library(tidyverse)
setwd('/Volumes/GoogleDrive/Mon disque/wrangling/project/wranglinghub')

data <- read_csv('football_data.csv')
```

```
## Warning: One or more parsing issues, call ‘problems()‘ on your data frame for details,
## e.g.:
##   dat <- vroom(...)
##   problems(dat)
```

```
## Rows: 13504 Columns: 36
## -- Column specification -------------------------------------------------------
## Delimiter: ","
## chr (17): schedule_date, schedule_week, team_home, Home team abbrev, team_aw...
## dbl (17): schedule_season, score_home, score_away, spread_favorite, over_und...
## lgl  (2): schedule_playoff, stadium_neutral
##
## 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.
```

```
head(data)
```

```
## # A tibble: 6 x 36
##   schedule_date schedu~1 sched~2 sched~3 team_~4 Home ~5 score~6 score~7 team_~8
##   <chr>            <dbl> <chr>   <lgl>   <chr>   <chr>     <dbl>   <dbl> <chr>
## 1 9/2/1966          1966 1       FALSE   Miami ~ MIA          14      23 Oaklan~
## 2 9/3/1966          1966 1       FALSE   Housto~ TEN          45       7 Denver~
## 3 9/4/1966          1966 1       FALSE   San Di~ LAC          27       7 Buffal~
## 4 9/9/1966          1966 2       FALSE   Miami ~ MIA          14      19 New Yo~
## 5 9/10/1966         1966 1       FALSE   Green ~ GB           24       3 Baltim~
## 6 9/10/1966         1966 2       FALSE   Housto~ TEN          31       0 Oaklan~
## # ... with 27 more variables: ‘Away team abbrev‘ <chr>, team_favorite_id <chr>,
## #   spread_favorite <dbl>, over_under_line <dbl>, stadium <chr>,
## #   stadium_neutral <lgl>, weather_temperature <dbl>, weather_wind_mph <dbl>,
## #   weather_humidity <dbl>, weather_detail <chr>,
## #   Difference_favored_minus_notfavored <dbl>, ‘Abs value of spread‘ <dbl>,
## #   ‘Actual difference - spread‘ <dbl>, stadium_location <chr>,
## #   stadium_open <dbl>, stadium_close <dbl>, stadium_type <chr>, ...
```

MISSING DATA

```
summary(data)
```

```
##  schedule_date      schedule_season schedule_week      schedule_playoff
##  Length:13504       Min.   :1966    Length:13504       Mode :logical
##  Class :character   1st Qu.:1983    Class :character   FALSE:12957
##  Mode  :character   Median :1997    Mode  :character   TRUE :547
##                     Mean   :1996
##                     3rd Qu.:2010
##                     Max.   :2022
##
##   team_home          Home team abbrev     score_home      score_away
##  Length:13504       Length:13504       Min.   : 0.00   Min.   : 0.00
##  Class :character   Class :character   1st Qu.:15.00   1st Qu.:13.00
##  Mode  :character   Mode  :character   Median :22.00   Median :20.00
##                                        Mean   :22.45   Mean   :19.76
##                                        3rd Qu.:29.00   3rd Qu.:27.00
##                                        Max.   :72.00   Max.   :62.00
##                                        NA's   :255     NA's   :255
##   team_away          Away team abbrev   team_favorite_id   spread_favorite
##  Length:13504       Length:13504       Length:13504       Min.   :-26.500
##  Class :character   Class :character   Class :character   1st Qu.: -7.000
##  Mode  :character   Mode  :character   Mode  :character   Median : -4.500
##                                                           Mean   : -5.394
##                                                            3rd Qu.: -3.000
##                                                            Max.   :  0.000
##                                                            NA’s   :2735
##  over_under_line   stadium          stadium_neutral weather_temperature
##  Min.   :28.00   Length:13504       Mode :logical   Min.   :-6.00
##  1st Qu.:38.50   Class :character   FALSE:13396     1st Qu.:48.00
##  Median :42.00   Mode  :character   TRUE :108       Median :62.00
##  Mean   :42.22                                      Mean   :58.87
##  3rd Qu.:45.00                                      3rd Qu.:72.00
##  Max.   :63.50                                      Max.   :97.00
##  NA’s   :2807                                       NA’s   :1224
##  weather_wind_mph weather_humidity weather_detail
##  Min.   : 0.000   Min.   :  4.00   Length:13504
##  1st Qu.: 3.000   1st Qu.: 57.00   Class :character
##  Median : 8.000   Median : 69.00   Mode  :character
##  Mean   : 7.689   Mean   : 67.22
##  3rd Qu.:11.000   3rd Qu.: 79.00
##  Max.   :40.000   Max.   :100.00
##  NA’s   :1240     NA’s   :5063
##  Difference_favored_minus_notfavored Abs value of spread
##  Min.   :-45.00                      Min.   : 0.000
##  1st Qu.: -3.00                      1st Qu.: 3.000
##  Median :  4.00                      Median : 4.500
##  Mean   :  5.26                      Mean   : 5.394
##  3rd Qu.: 14.00                      3rd Qu.: 7.000
##  Max.   : 59.00                      Max.   :26.500
##  NA’s   :2735                        NA’s   :2735
##  Actual difference - spread stadium_location    stadium_open  stadium_close
##  Min.   :-52.0000           Length:13504       Min.   :1909   Min.   :1970
##  1st Qu.: -8.5000           Class :character   1st Qu.:1966   1st Qu.:1996
##  Median : -0.5000           Mode  :character   Median :1975   Median :2001
##  Mean   : -0.1342                              Mean   :1978   Mean   :2003
##  3rd Qu.:  8.5000                              3rd Qu.:1998   3rd Qu.:2009
##  Max.   : 50.5000                              Max.   :2020   Max.   :2016
##  NA’s   :2735                                  NA’s   :554    NA’s   :7231
##  stadium_type       stadium_address    stadium_weather_station_code
##  Length:13504       Length:13504       Length:13504
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  stadium_weather_type stadium_capacity stadium_surface      STATION
##  Length:13504         Min.   :27000    Length:13504       Length:13504
##  Class :character     1st Qu.:65890    Class :character   Class :character
##  Mode  :character     Median :71250    Mode  :character   Mode  :character
##                       Mean   :71100
##                       3rd Qu.:76416
##                       Max.   :93605
##                       NA’s   :6307
##      NAME              LATITUDE       LONGITUDE         ELEVATION
##  Length:13504       Min.   :25.79   Min.   :-122.41   Min.   :   1.8
##  Class :character   1st Qu.:33.94   1st Qu.: -97.07   1st Qu.:   8.8
##  Mode  :character   Median :39.10   Median : -84.52   Median : 145.4  
##                     Mean   :37.89   Mean   : -90.29   Mean   : 188.0  
##                     3rd Qu.:40.68   3rd Qu.: -78.89   3rd Qu.: 209.4  
##                     Max.   :47.65   Max.   : -71.00   Max.   :1611.2  
##                     NA's   :3220    NA's   :3220      NA's   :3220    
```

Visualizing which columns have missing data. "weather_detail" has the most missing values by far. This variable contains 8 factors: "DOME", "DOME (Open Roof)", "Fog", "Rain", "Rain, Fog", "Snow", "Snow, Fog", and "Snow, Freezing Rain". All of these factors are related to poor weather, or dome-related changes. Values were not recorded for "nice" weather, which is probably contributing to the large amount of missing values. To proceed, we simply decided not to use this column in our analysis. There were many other weather-related columns available in the dataset.

```
# data_missing <- data
# library(naniar)
# gg_miss_var(data_missing)

# levels(as.factor(data_missing$weather_detail))
```

```
# library(naniar)
# vis_miss(data_missing)
```

```
library(naniar)

# mcar_test(data.frame(data)) # this runs into an error, but we can subset the data by columns and
# run an mcar test on just a portion of the data

# this confirms that the data is missing not a random
mcar_test(data.frame(data[,10:20]))
```

```
## # A tibble: 1 x 4
##   statistic    df p.value missing.patterns
##       <dbl> <dbl>   <dbl>            <int>
## 1    10578.   102       0               15
```

```
# library(naniar)
#
# gg_miss_fct(x = data_missing, fct = schedule_season)
# a lot of missing data related to betting up until near 1978
# some missing data in the 2023 season, where the authors of the dataset pre-filled data for some games
```

Evidently, there is a pattern in our missing data. We have 100% missing data in the earlier football seasons. This makes sense because when football first emerged, betting was not yet available.

```
pct_miss(data)
```

```
## [1] 14.73679
```

```
betting_columns <- c("team_favorite_id",
                     "spread_favorite",
                     "over_under_line",
                     "Difference_favored_minus_notfavored",
                     "Actual difference - spread",
                     "Abs value of spread")
pct_miss(data[,betting_columns]) # 20.34%
```

```
## [1] 20.34212
```

```
pct_miss(data[, -which(colnames(data) %in% betting_columns)])
```

```
## [1] 13.61572
```

Looking at the data in excel, our regular season betting data begins at row 2494 (2493 if we do not include the header).

```
pct_miss(data[1:2492,betting_columns])
```

```
## [1] 99.49171
```

```
pct_miss(data[2493:13248,betting_columns])
```

```
## [1] 0.1084666
```

```
pct_miss(data[13249:nrow(data),betting_columns])
```

```
## [1] 100
```

WEATHER

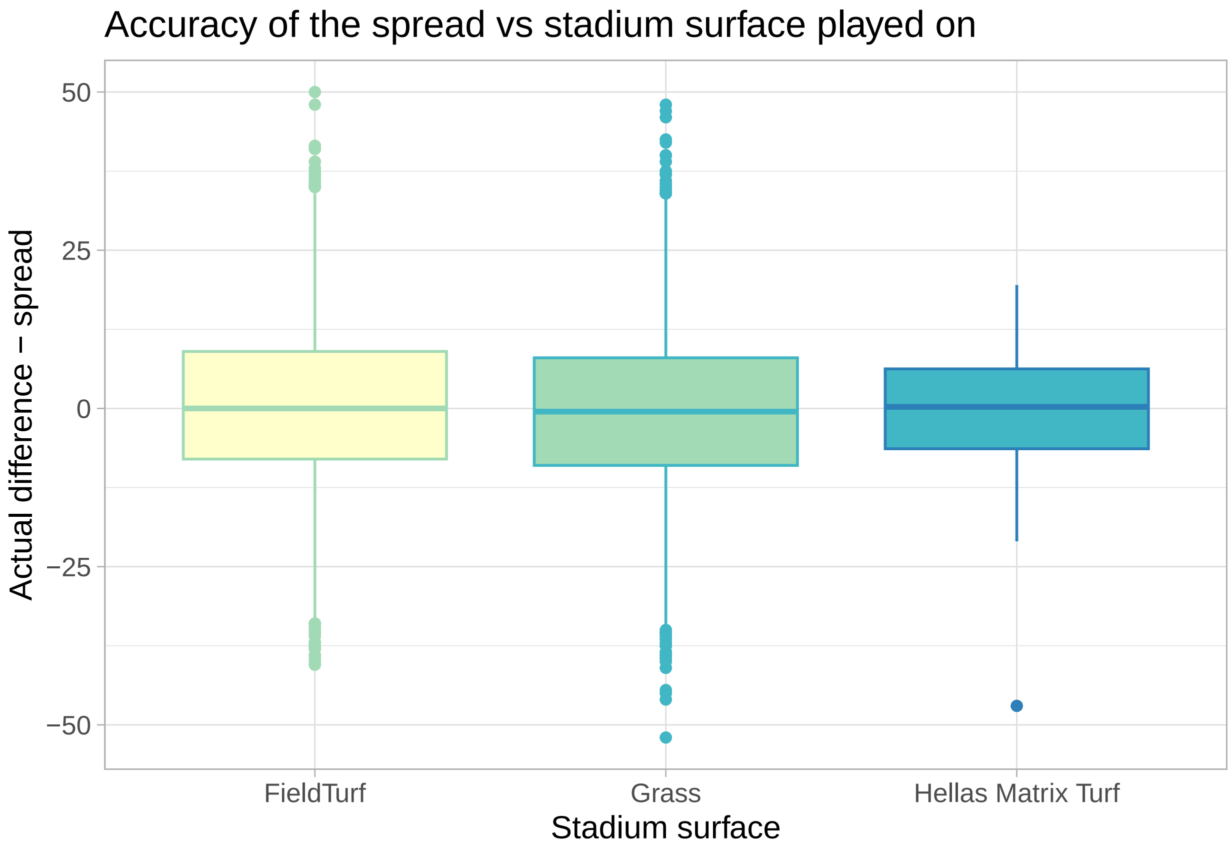
```
# weather_temperature
# weather_wind_mph
# weather_humidity

# stadium_weather_type
# stadium_surface

# Actual difference - spread

betting_data <- data[2493:13248,]
```

Playing surface:

```
table(as.factor(betting_data$stadium_surface))
```

```
##
##          FieldTurf              Grass Hellas Matrix Turf
##               2327               4016                 38
```

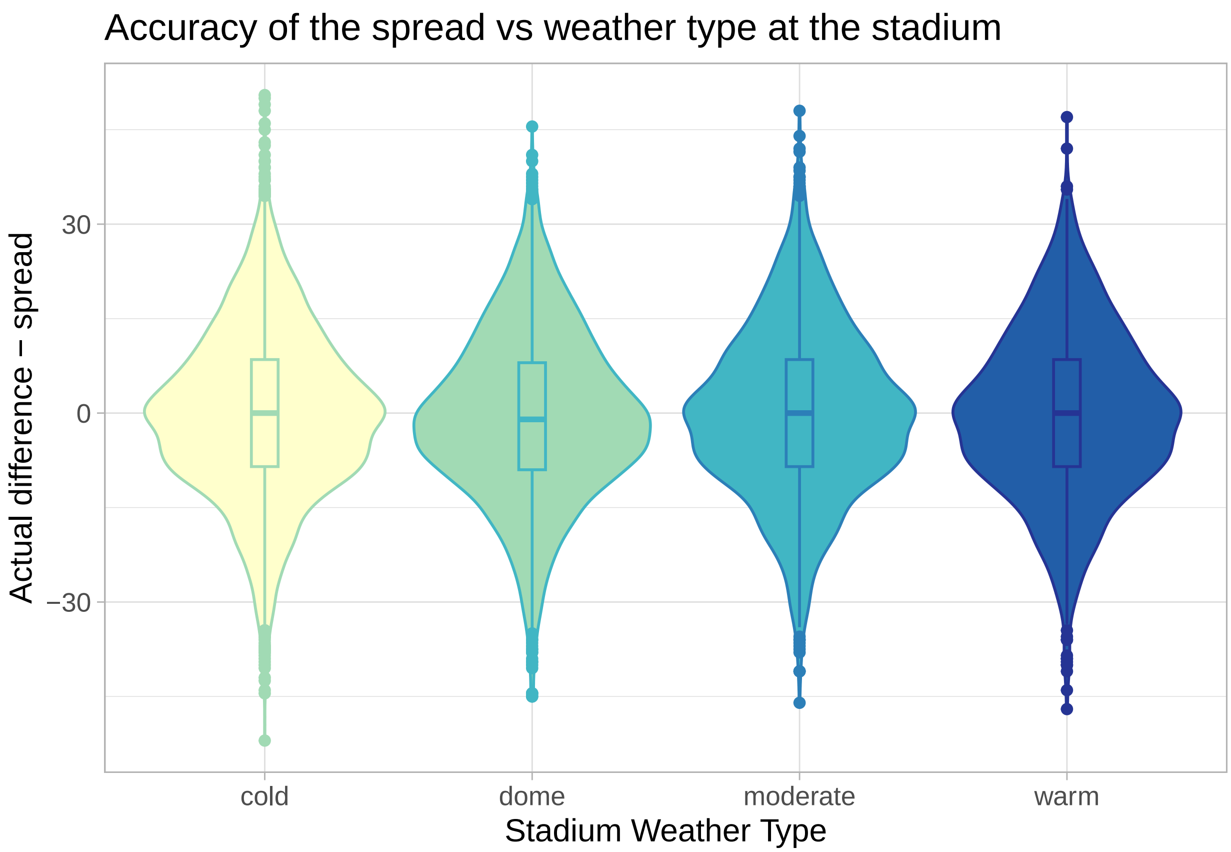
```
plot_playing_surface(betting_data)
```



```
table(betting_data$stadium_weather_type)
```

```
##
##     cold     dome moderate     warm
##     4349     2584     2098     1711
```

```
plot_stadium_weather(betting_data)
```



Temperature

```
# weather_temperature
# weather_wind_mph
# weather_humidity
domed_stadiums <- filter(betting_data,
                       stadium_weather_type=='dome')

non_domed_stadiums <- filter(betting_data,
                       stadium_weather_type=='cold' |
                         stadium_weather_type=='moderate' |
                         stadium_weather_type=='warm')

summary(non_domed_stadiums$weather_temperature) # 50,64,72
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   -6.00   45.00   57.00   55.83   67.50   97.00     799
```

```
summary(non_domed_stadiums$weather_wind_mph) # 0, 7, 11
```

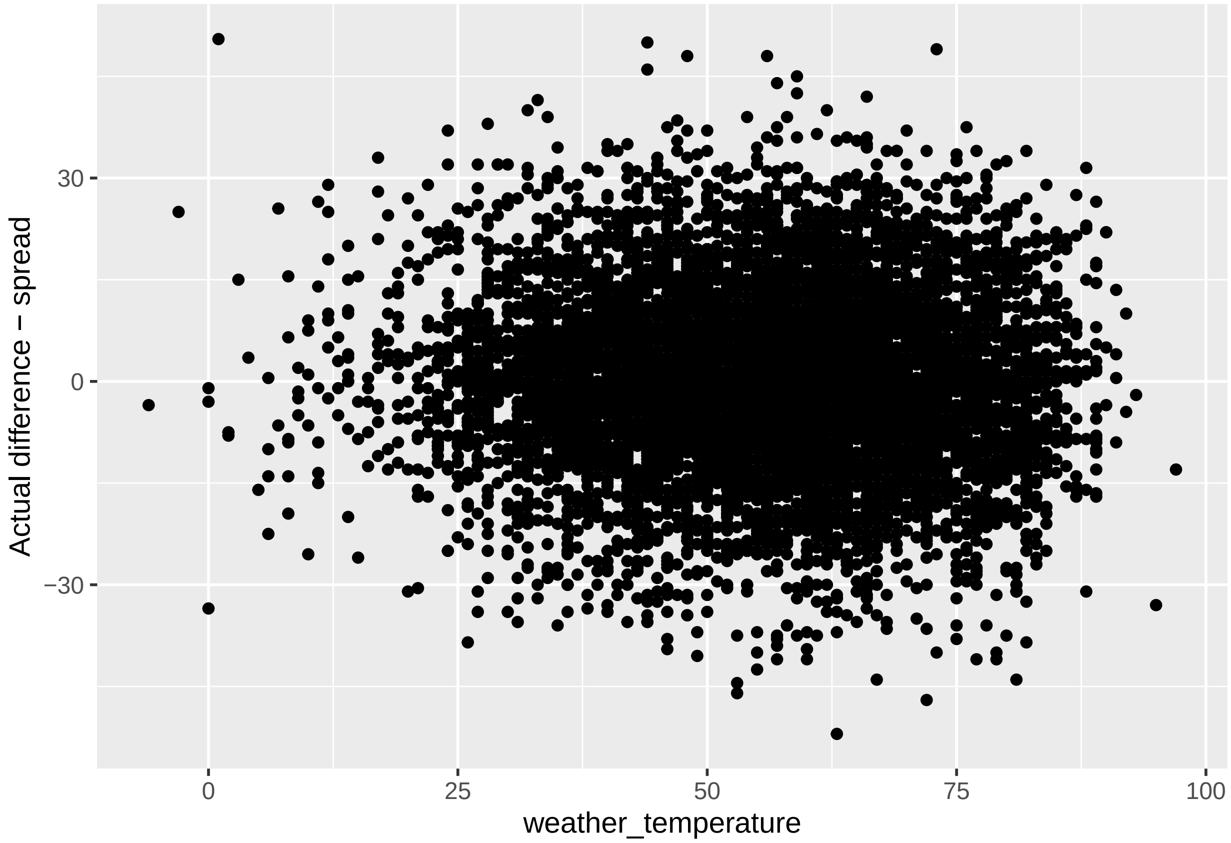
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   6.000   9.000   9.538  12.000  40.000     810
```

```
summary(non_domed_stadiums$weather_humidity) # 56, 68, 78
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    4.00   57.00   69.00   67.03   78.00  100.00    1921
```

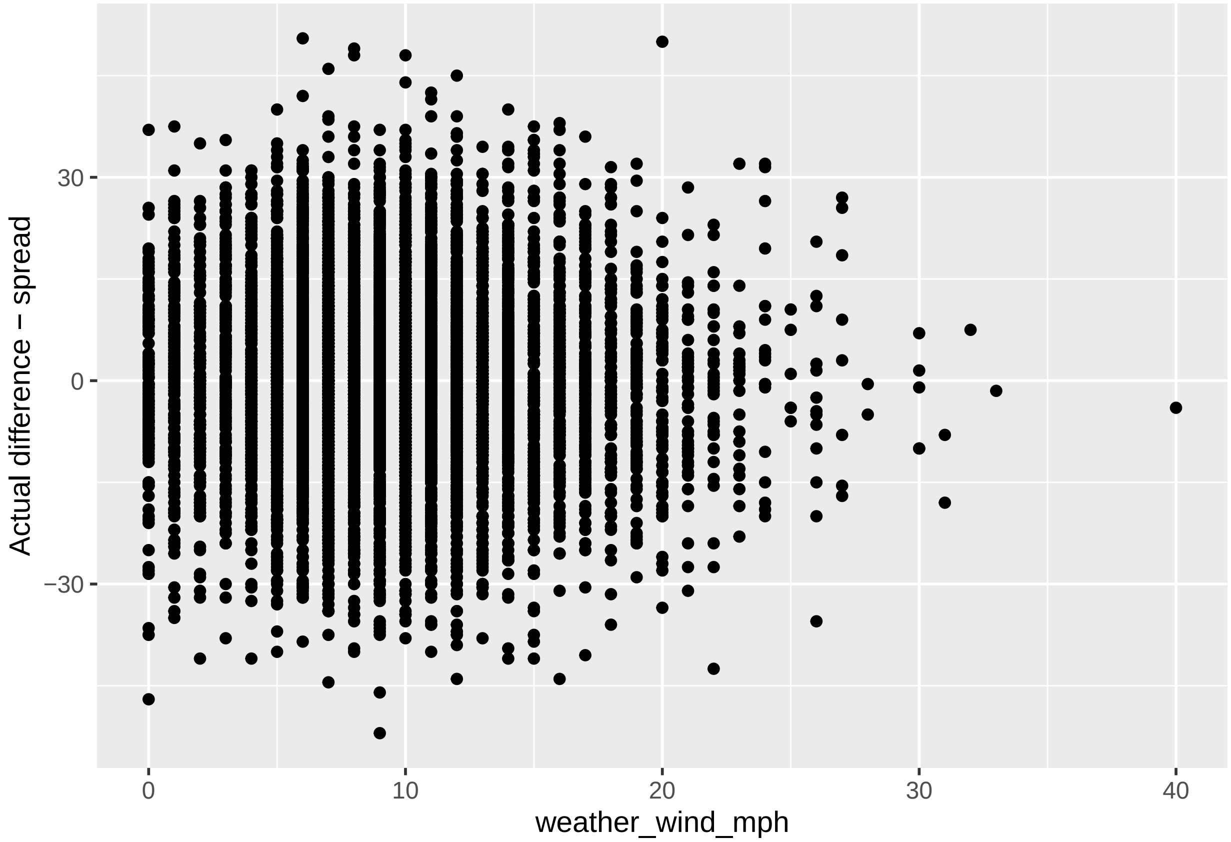
```
# new values (when removing domed stadiums):
# 45, 57, 67.5
# 6, 9, 12
# 57, 69, 78

ggplot(non_domed_stadiums, aes(x=weather_temperature, y=`Actual difference - spread`)) +
  geom_point()
```

```
## Warning: Removed 799 rows containing missing values (`geom_point()`).
```



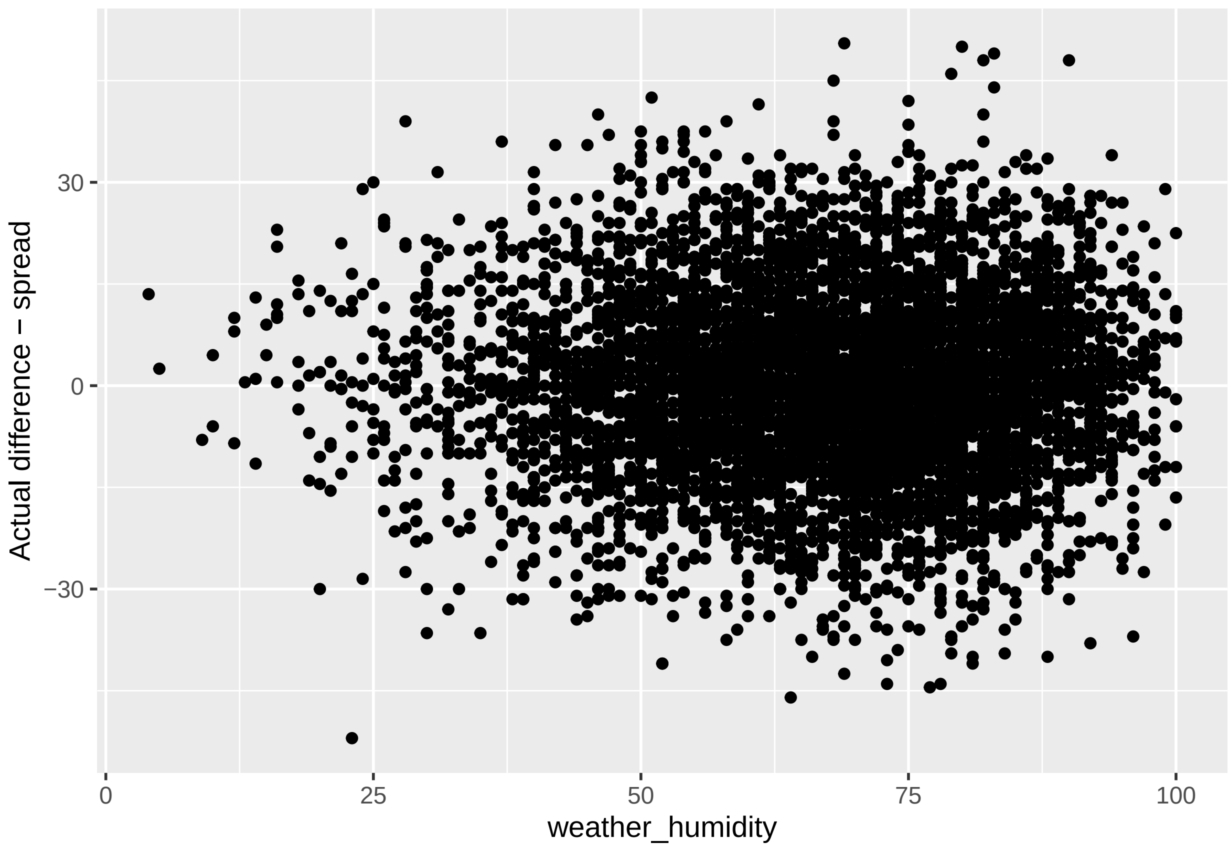
```
ggplot(non_domed_stadiums, aes(x=weather_wind_mph, y=`Actual difference - spread`)) +
  geom_point()
```

```
## Warning: Removed 810 rows containing missing values (`geom_point()`).
```



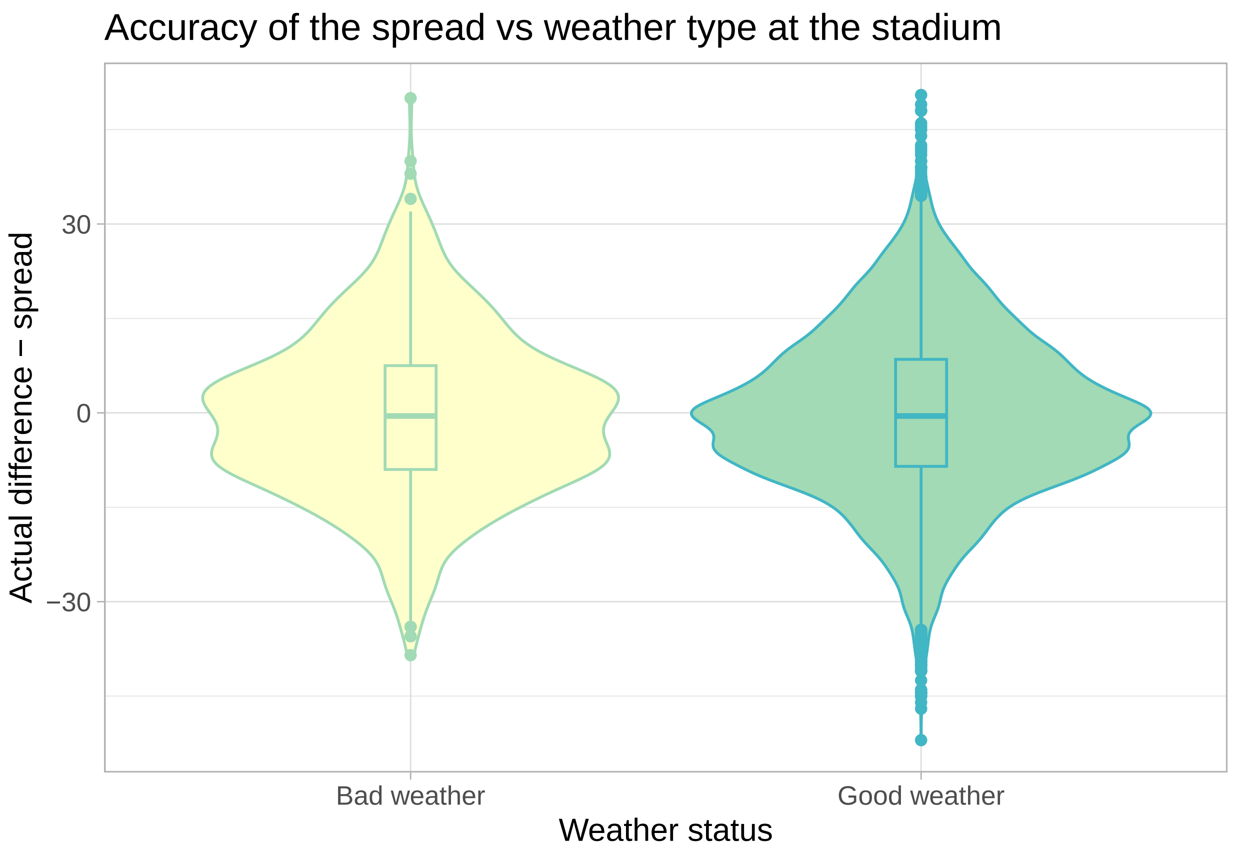
```
ggplot(non_domed_stadiums, aes(x=weather_humidity, y=`Actual difference - spread`)) +
  geom_point()
```

```
## Warning: Removed 1921 rows containing missing values (‘geom_point()‘).
```



Categorical weather:

```
plot_weather_status(betting_data)
```


Accuracy of the spread vs weather type at the stadium
Actual difference – spread
30
0
−30
Bad weather
Good weather
Weather status