strava_analysis

Frank Neugebauer

10/11/2019

Working with Strava Data in R

The R work is pretty straightfoward. This R Markdown file uses the `all_data.xlsx` file created in the `get_strava_data.r` file. This must be created before anything within this Jupyter Notebook will work and you can't build that file using Jupyter Notebook (unless you happen to know how to capture the authorize step).

Note that the analysis focuses on cycling data, a fact that will manifest itself soon enough.

First things first - load the data by first create a data structure to specify the data types and then using that to open the `all_data.xlsx` file.

```
# All data
data_types <- c('character', 'numeric', 'numeric', 'numeric', 'numeric', 'numeric',
                'numeric', 'character', 'character', 'character', 'numeric',
                'numeric', 'numeric', 'numeric', 'numeric', 'numeric')

strava_data <- read.xlsx("./all_data.xlsx", 1, colClasses = data_types, header=TRUE)
```

Subset the Data - Part I

Many of my own Strava activites are not cycling (e.g., running, swimming). Furthermore, some of the cycling activites were on stationary bikes without trackable power meters, which means important data is missing for those activites. The next step is to only include relevant cycling activities.

```
# Only ride data
ride_data <- subset(strava_data, (type == 'VirtualRide' | type == 'Ride') &
                      average_watts > 0)
```

Some Descriptive Analytics

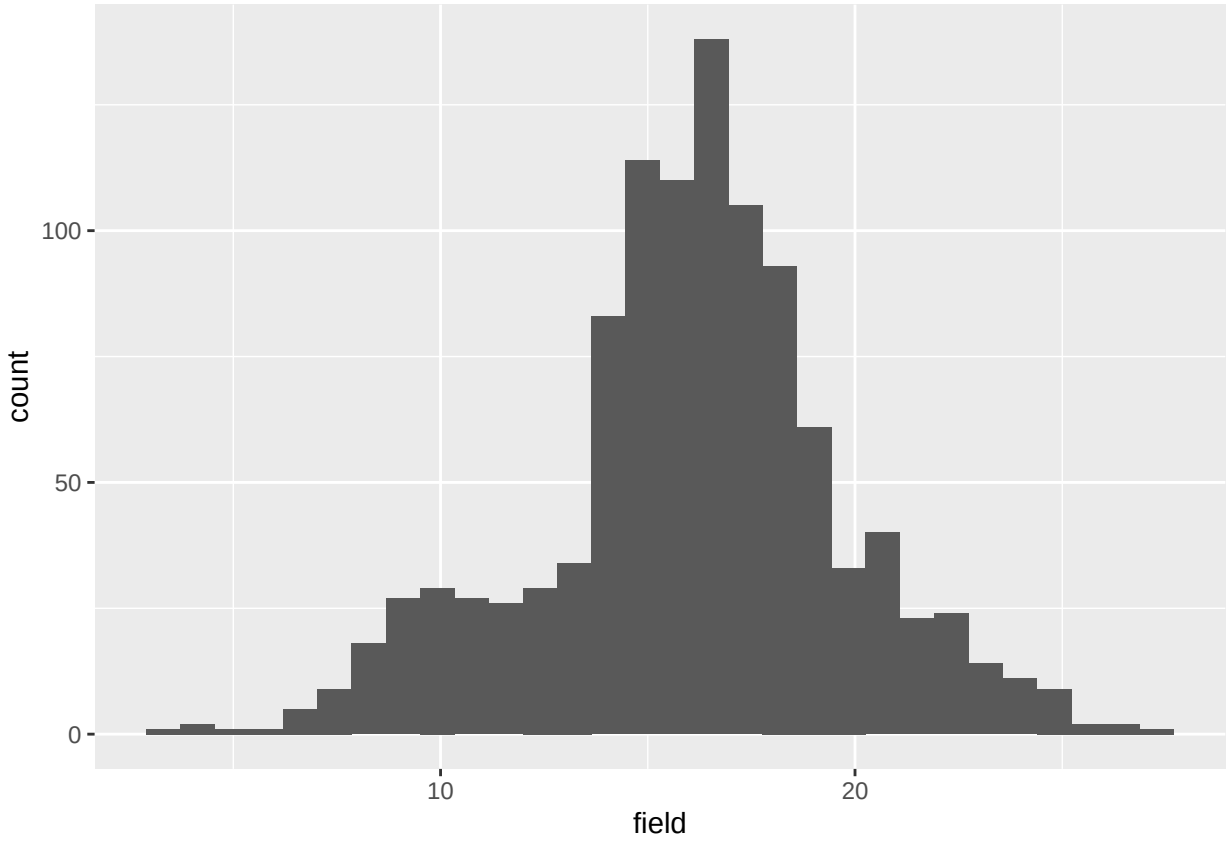
The following is a reusable function that outputs some important descriptive analytics, including:

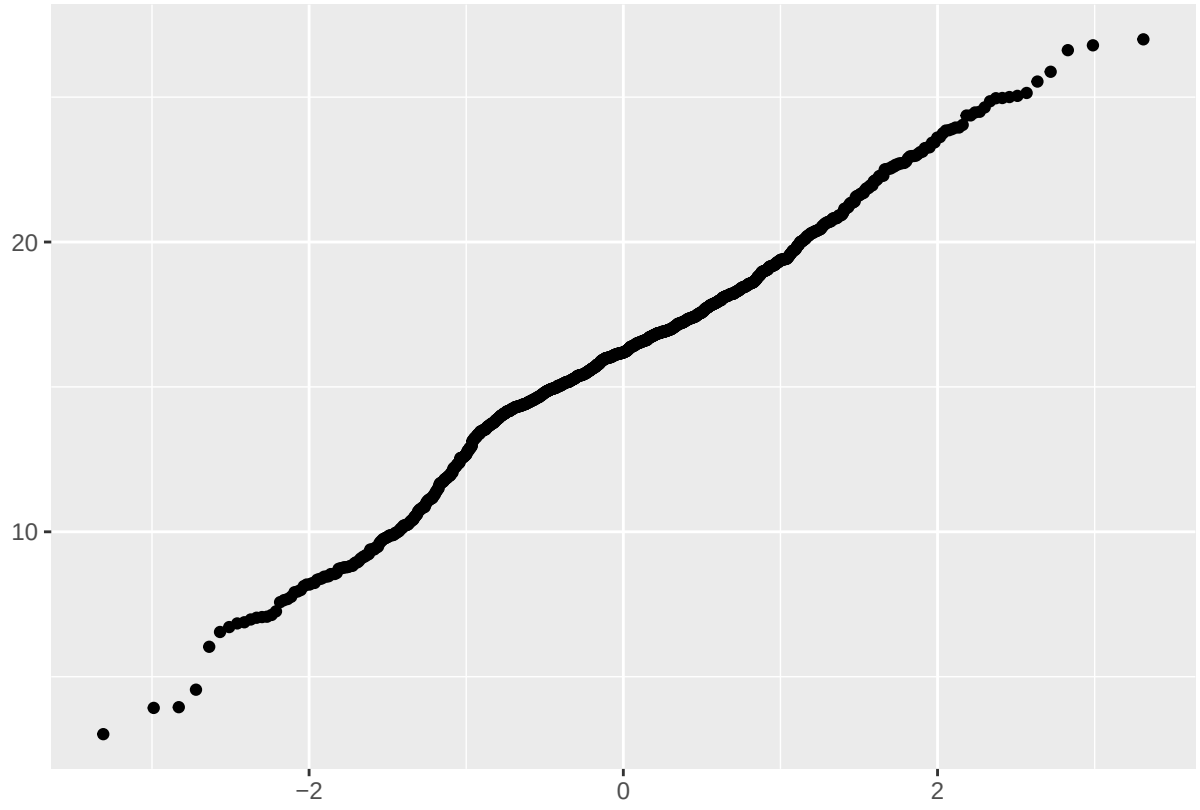
- histogram
- qplot
- a general description with mean, median, max, min, etc.
- a simple kurtosis analysis
- skew analysis

Call the function on key data elements.

```
a <- perf_analysis('Average Speed (MPH)', ride_data, ride_data$average_speed_mph, 15)
```

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

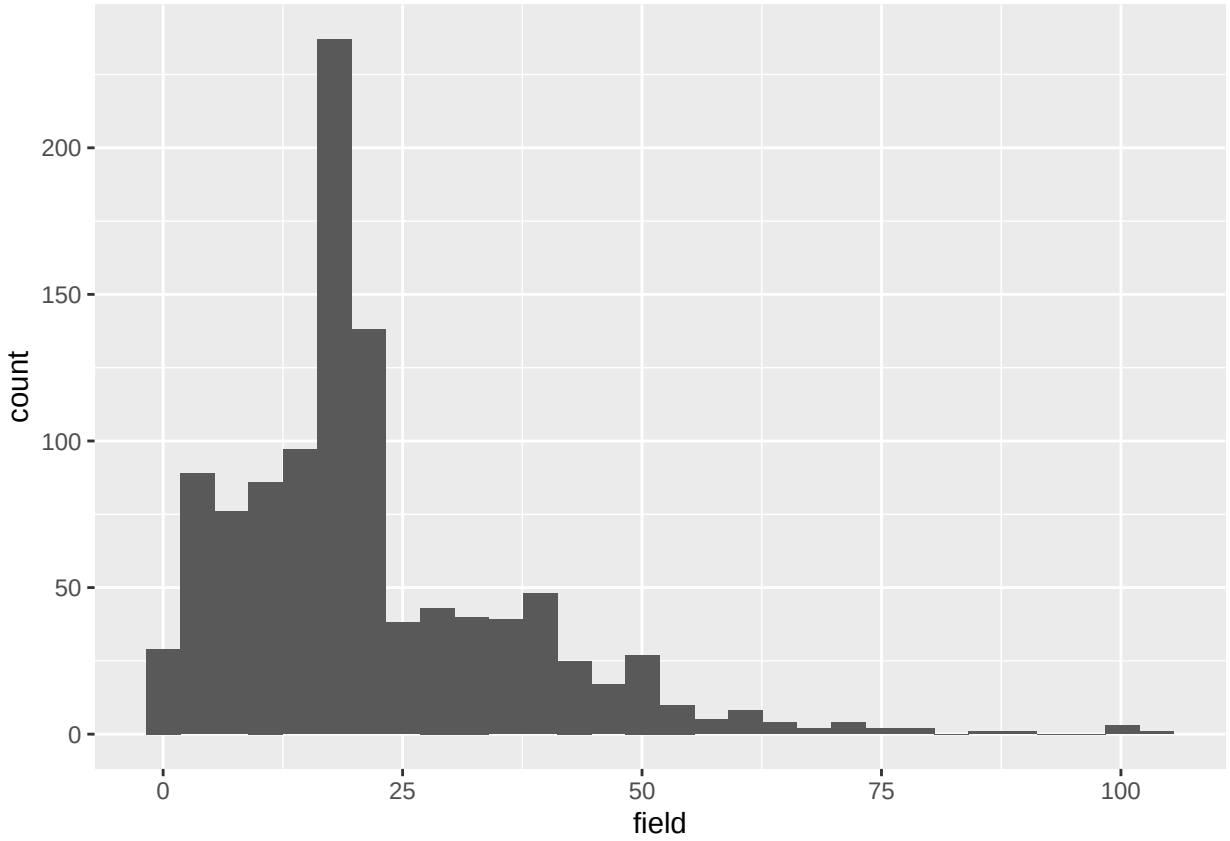

count
100
50
0
10
20
field

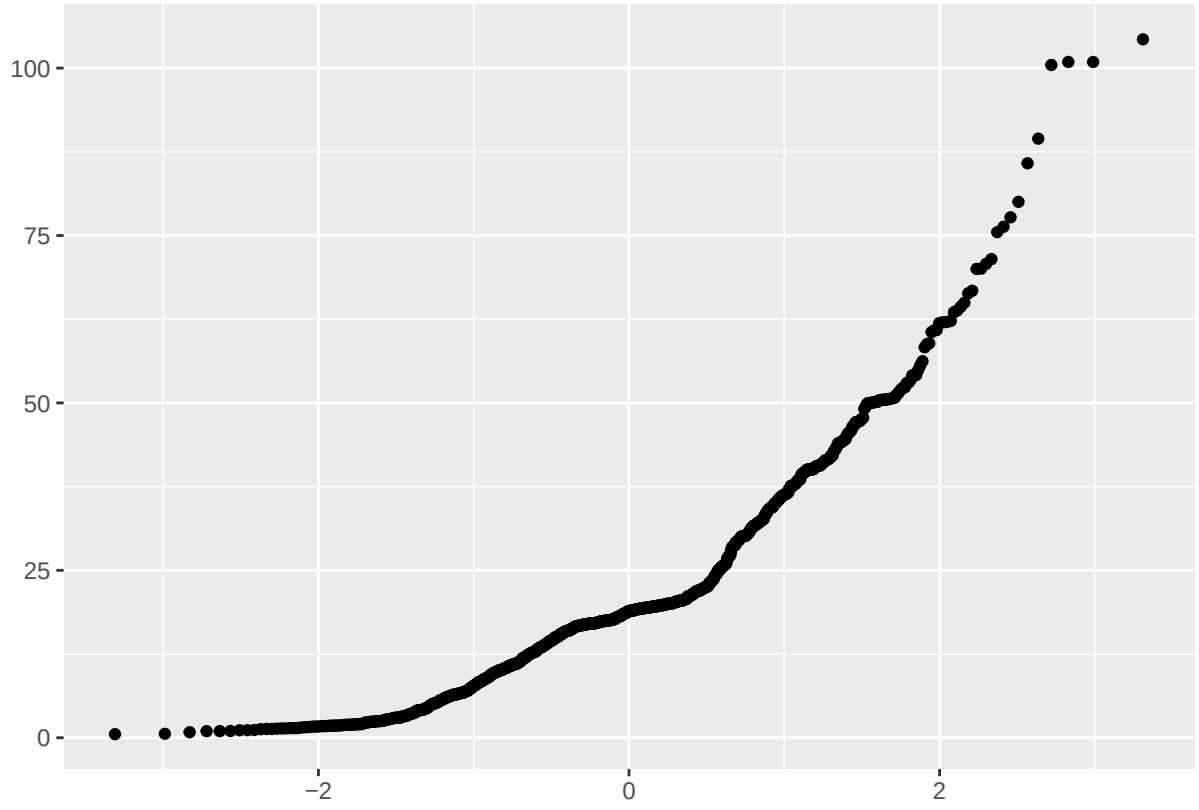



```
##    vars    n  mean   sd median trimmed  mad  min   max range  skew
## X1    1 1072 16.07 3.67  16.19   16.16 2.87 3.02 26.99 23.98 -0.22
##    kurtosis   se
## X1     0.48 0.11
## [1] "Kurtosis is 0.48 . Since it's greater than zero, there may be a heavily-tailed distribution. Id
## [1] "Skew is -0.22 . Since it's less than zero, there may be  a pile up of scores on the right of th
```

```
a <- perf_analysis('Distance (Miles)', ride_data, ride_data$distance_mi, 15)
```

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

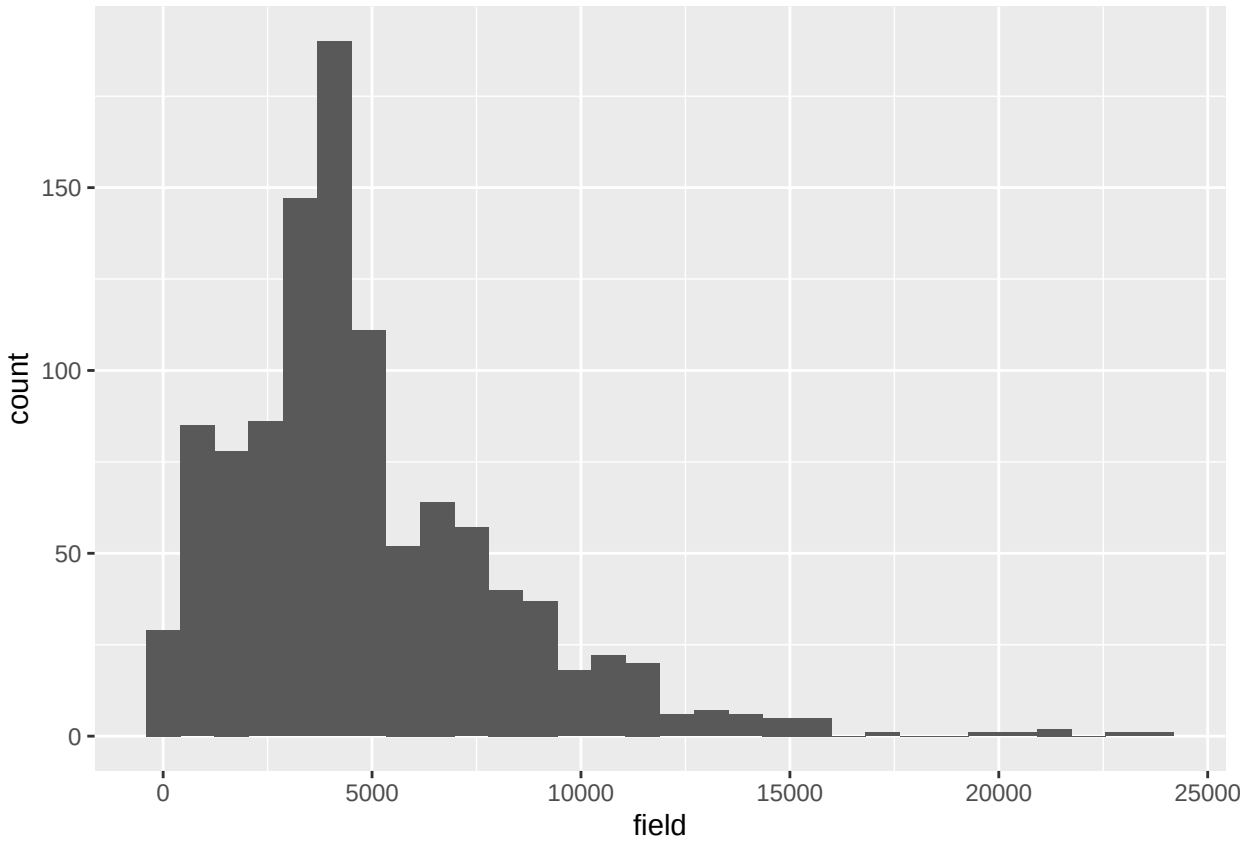

count
200
150
100
50
0
0
25
50
75
100
field

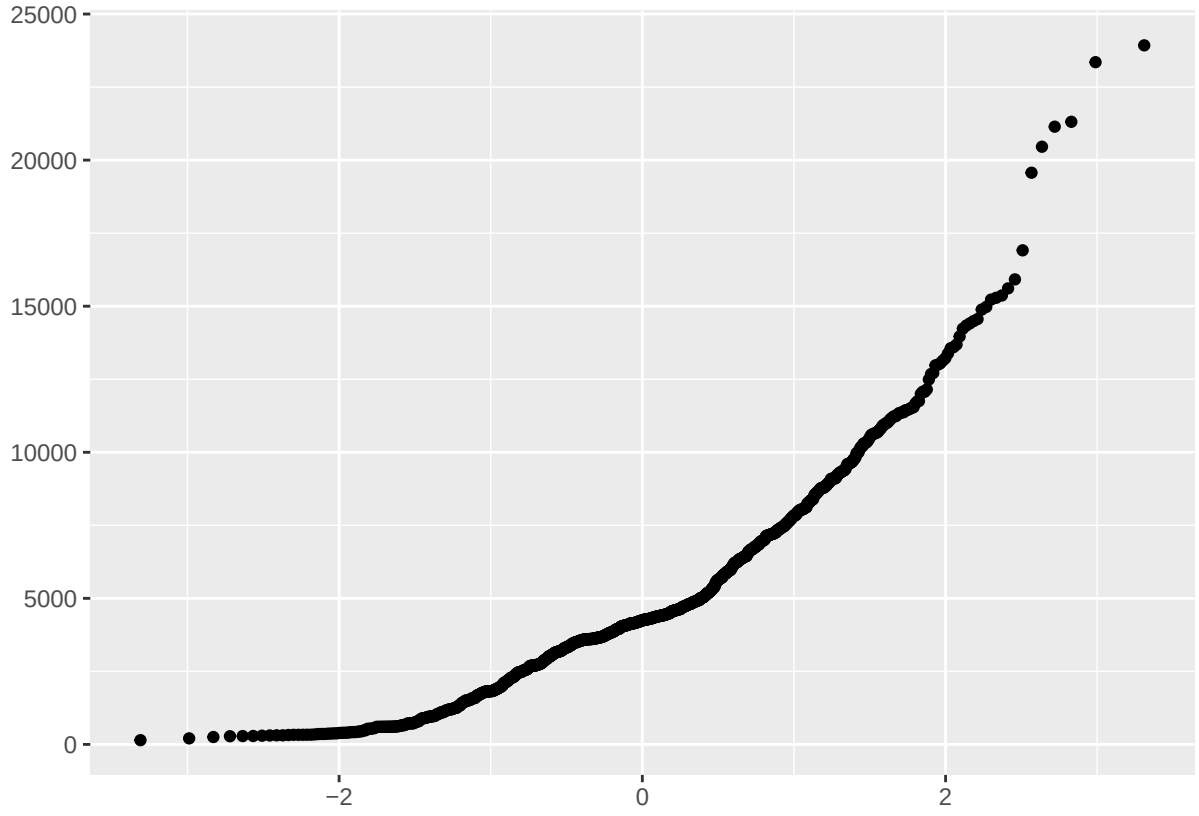



```
##    vars    n  mean    sd median trimmed   mad  min    max  range skew
## X1    1 1072 21.65 15.23  18.92   19.86 11.54 0.54 104.29 103.75 1.51
##    kurtosis   se
## X1     3.61 0.47
## [1] "Kurtosis is 3.61 . Since it's greater than zero, there may be a heavily-tailed distribution. Id
## [1] "Skew is 1.51 . Since it's greater than zero, there may be  a pile up of scores on the left of t
```

```
a <- perf_analysis('Moving Time (Minutes)', ride_data, ride_data$moving_time, 15)
```

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

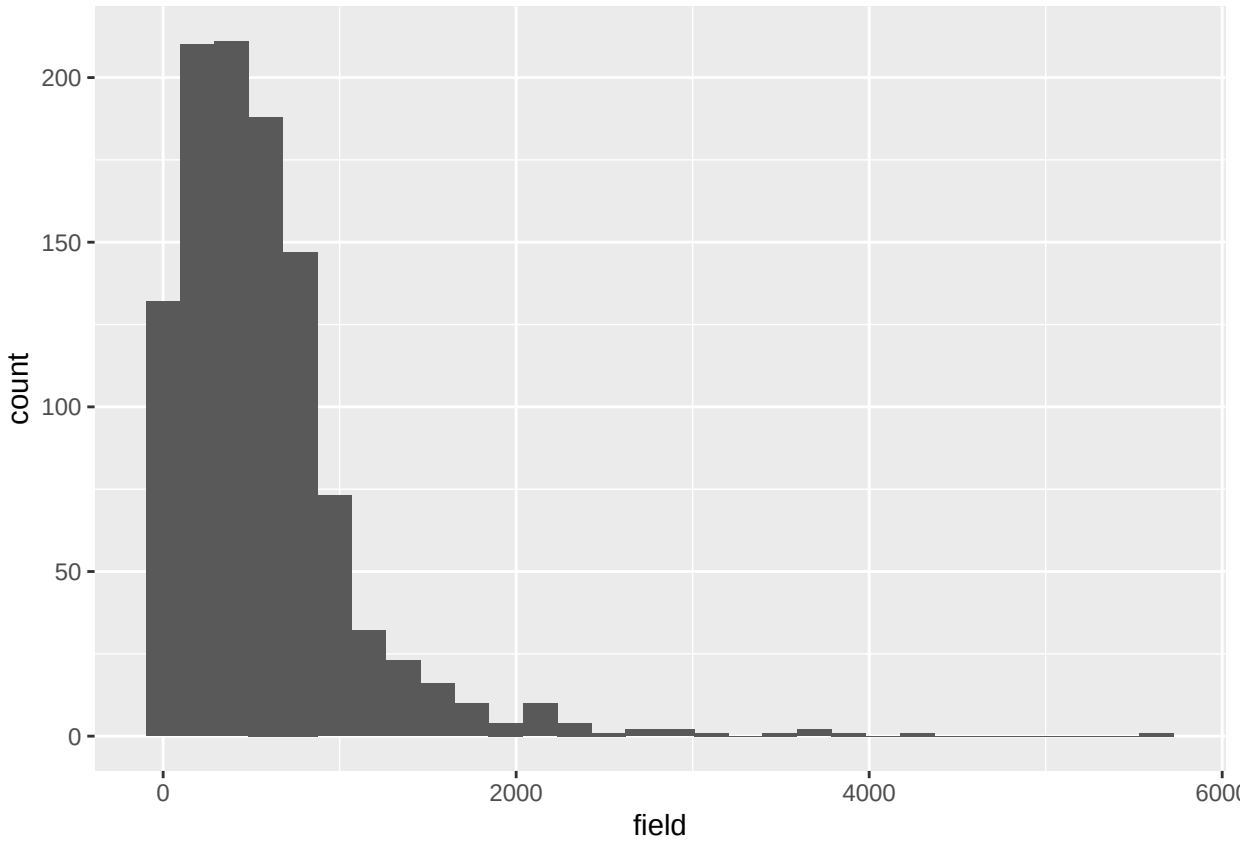

150
100
50
0
count
0
5000
10000
15000
20000
25000
field

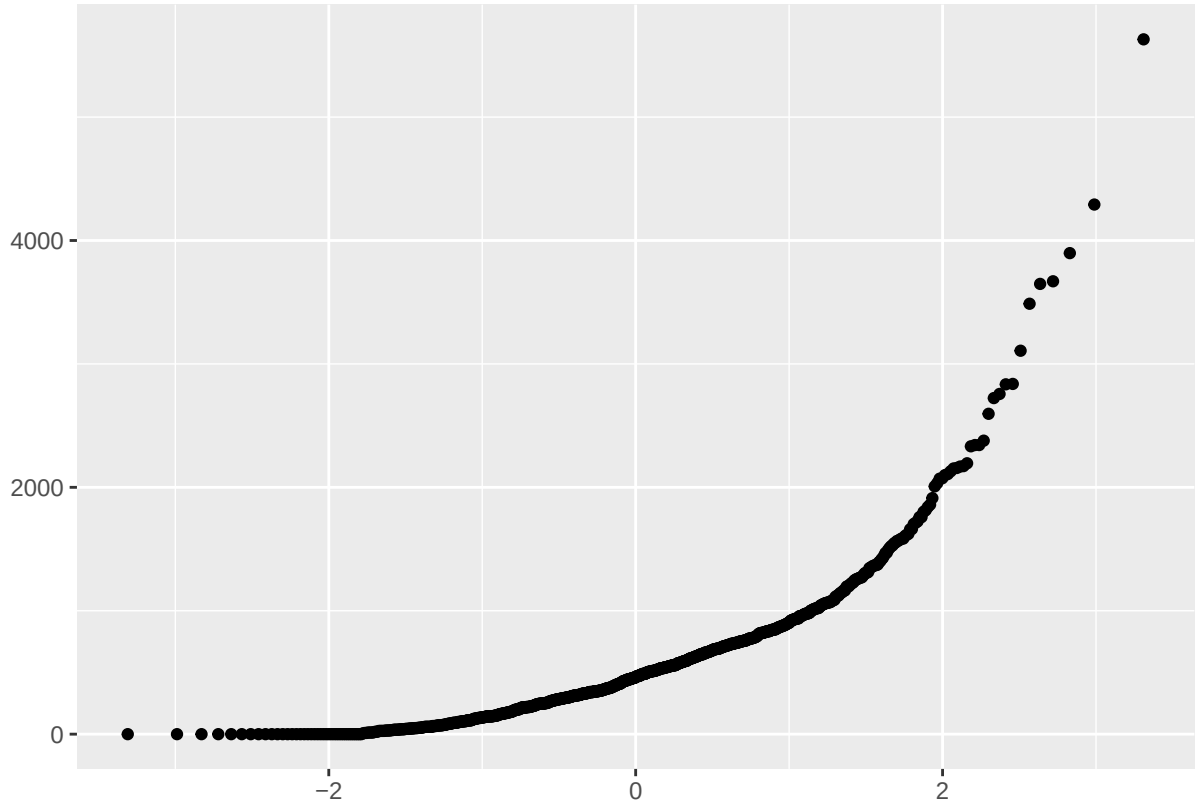



```
##    vars    n    mean      sd median trimmed     mad min   max range skew
## X1    1 1072 4880.99 3287.64   4251 4516.64 2541.18 146 23929 23783 1.48
##    kurtosis     se
## X1     3.83 100.41
## [1] "Kurtosis is 3.83 . Since it's greater than zero, there may be a heavily-tailed distribution. Id
## [1] "Skew is 1.48 . Since it's greater than zero, there may be  a pile up of scores on the left of t
```

```
a <- perf_analysis('Elevation Gain (Feet)', ride_data, ride_data$elevation_gain_ft, 15)
```

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

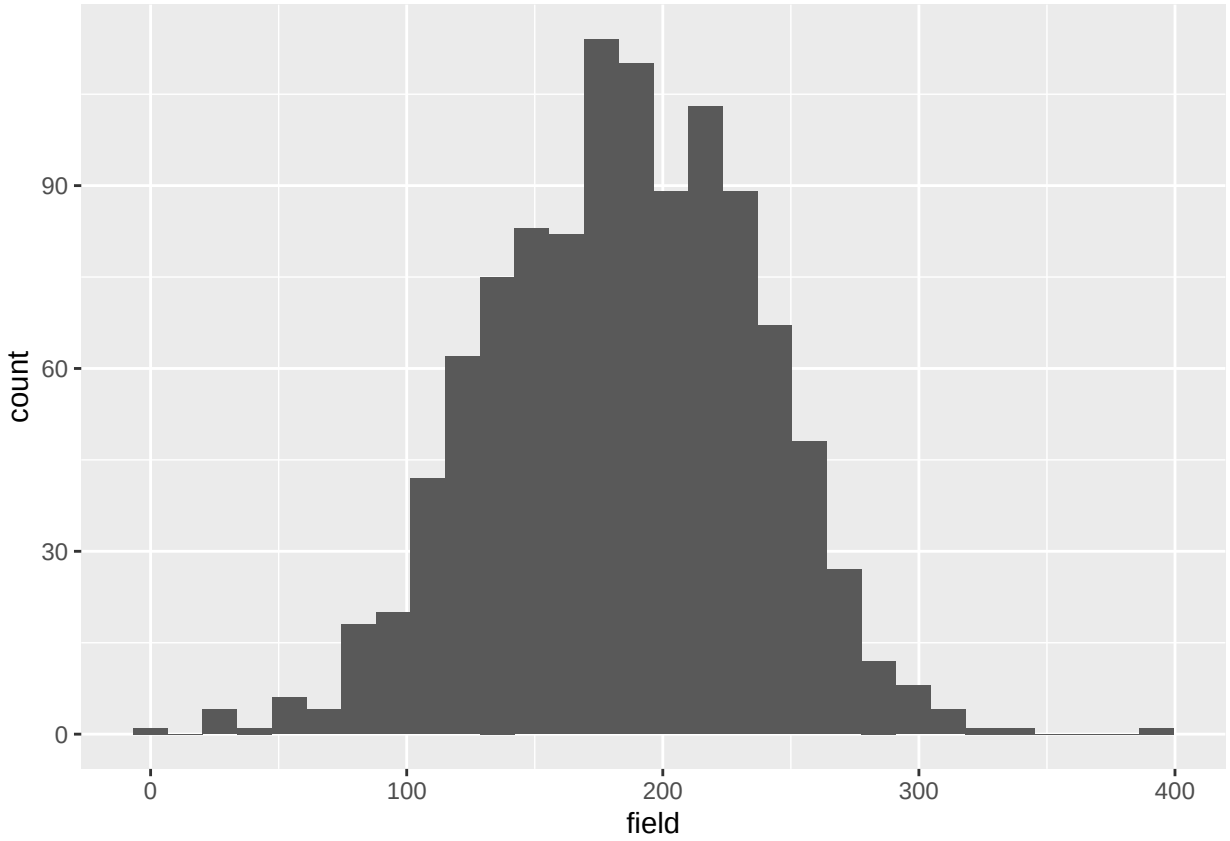

200
150
100
50
0
count
0
2000
4000
field

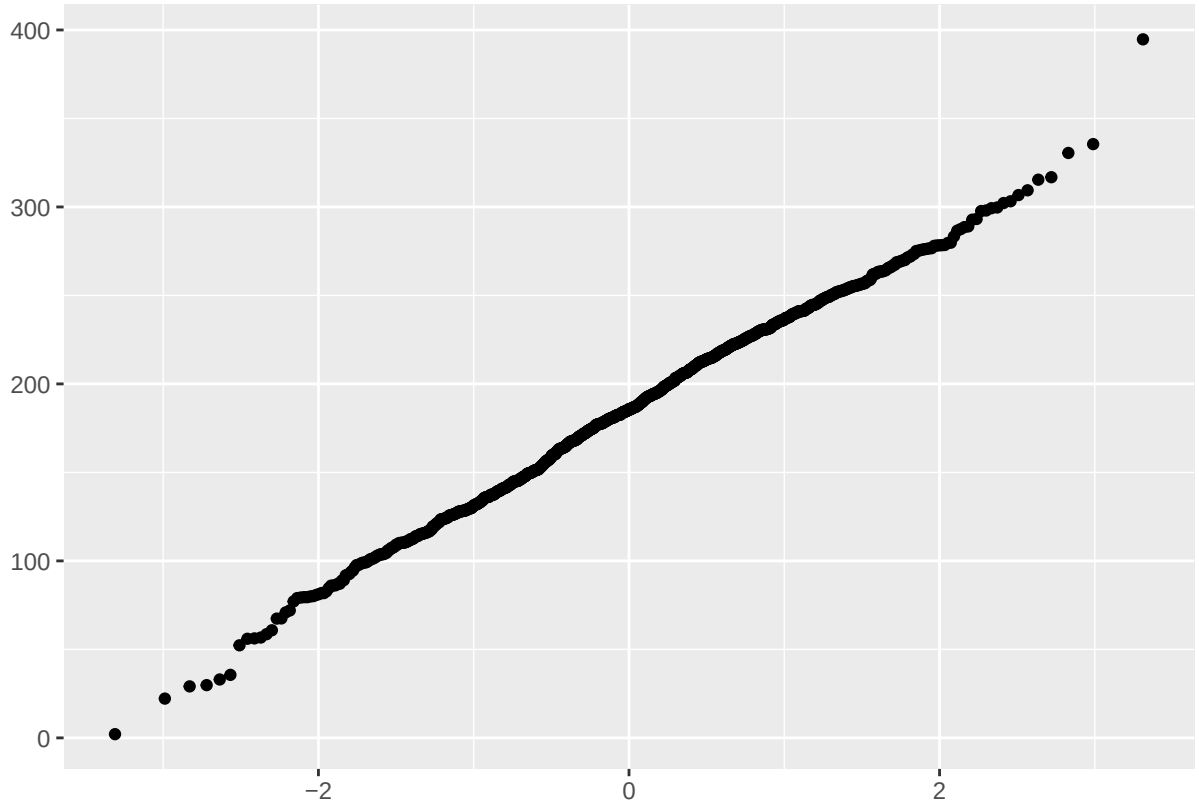



```
##    vars    n   mean     sd median trimmed    mad min     max   range skew
## X1    1 1072 565.67 531.12  462.6  486.78 381.84   0 5629.92 5629.92 2.91
##    kurtosis    se
## X1    15.38 16.22
## [1] "Kurtosis is 15.38 . Since it's greater than zero, there may be a heavily-tailed distribution. I
## [1] "Skew is 2.91 . Since it's greater than zero, there may be  a pile up of scores on the left of t
```

```
a <- perf_analysis('Average Power (Watts)', ride_data, ride_data$average_watts, 15)
```

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


count
90
60
30
0
0
100
200
300
400
field

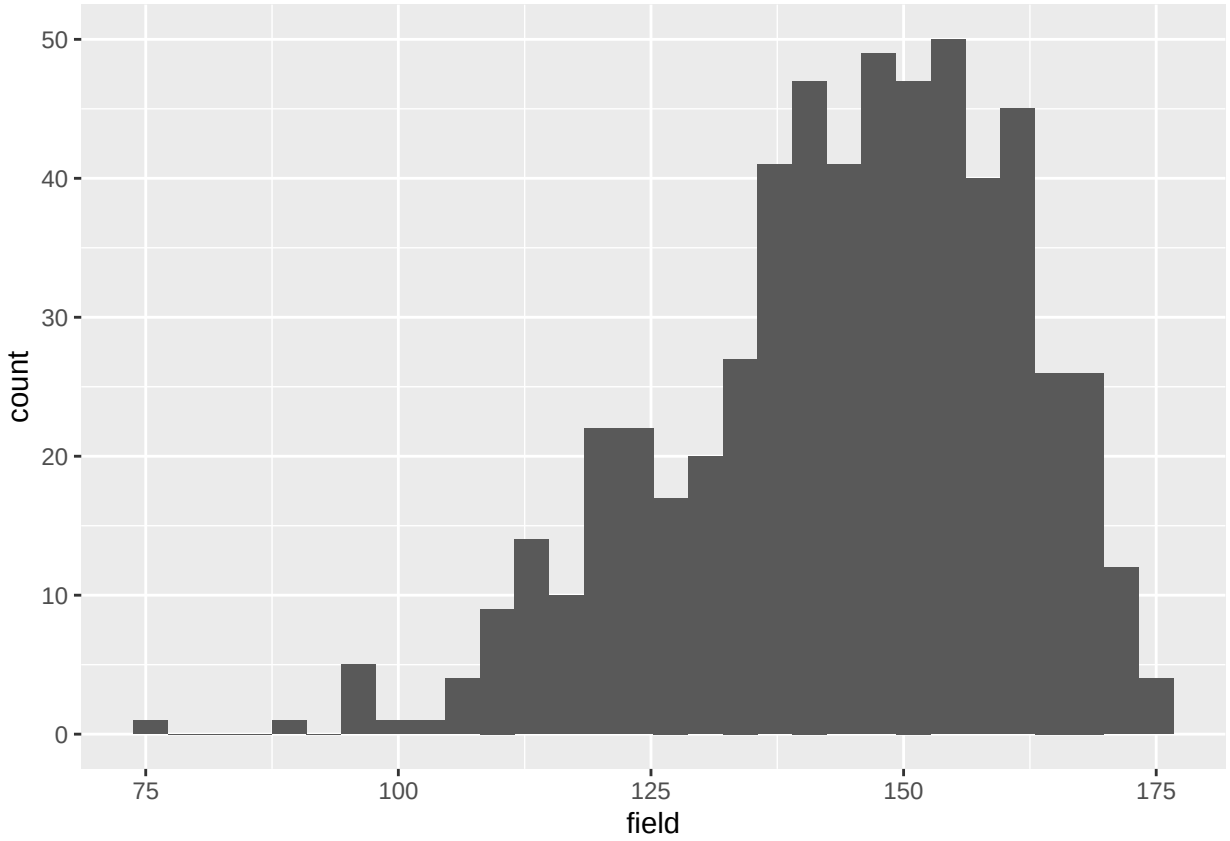



```
##    vars    n   mean    sd median trimmed   mad min   max range  skew
## X1    1 1072 184.83 51.44  185.5  185.51 55.15 2.1 394.7 392.6 -0.11
##    kurtosis   se
## X1     0.06 1.57
## [1] "Kurtosis is 0.06 . Since it's greater than zero, there may be a heavily-tailed distribution. Id
## [1] "Skew is -0.11 . Since it's less than zero, there may be  a pile up of scores on the right of th
```

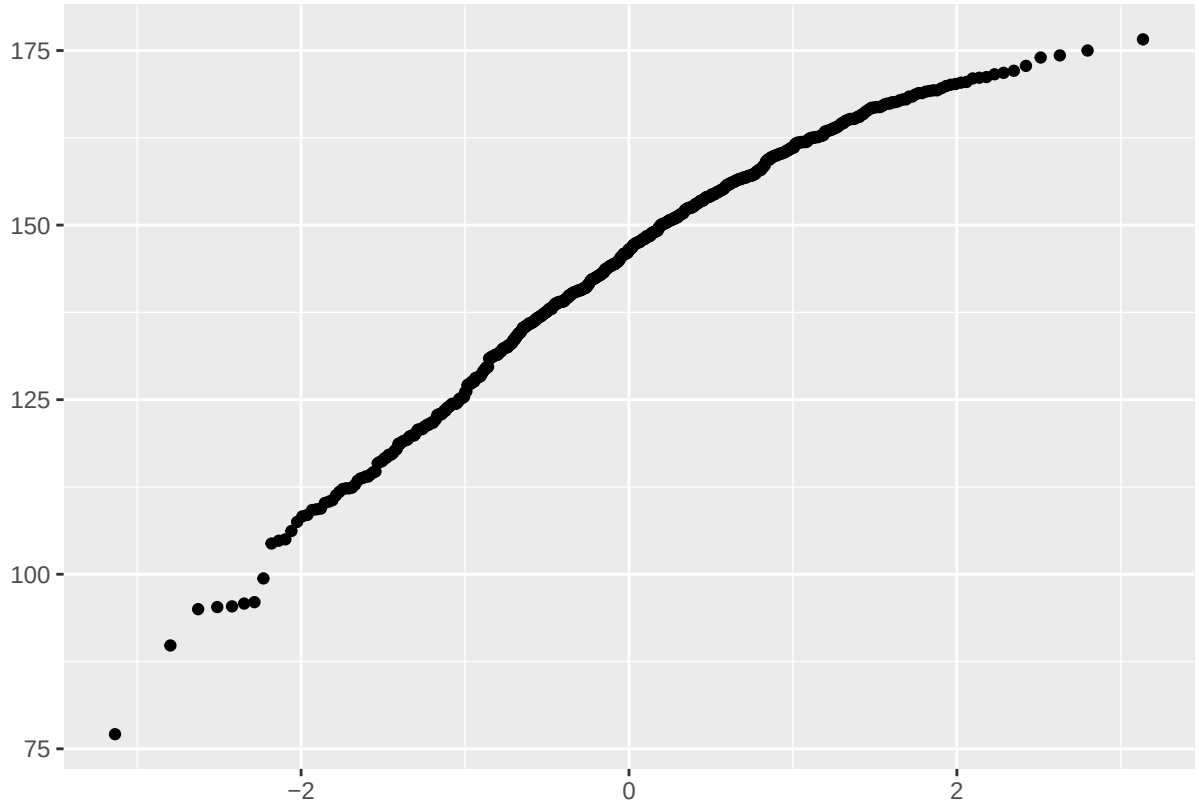
```
a <- perf_analysis('Average Heart Rate', ride_data, ride_data$average_heartrate, 15)
```

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

```
## Warning: Removed 490 rows containing non-finite values (stat_bin).
```




```
## Warning: Removed 490 rows containing non-finite values (stat_qq).
```

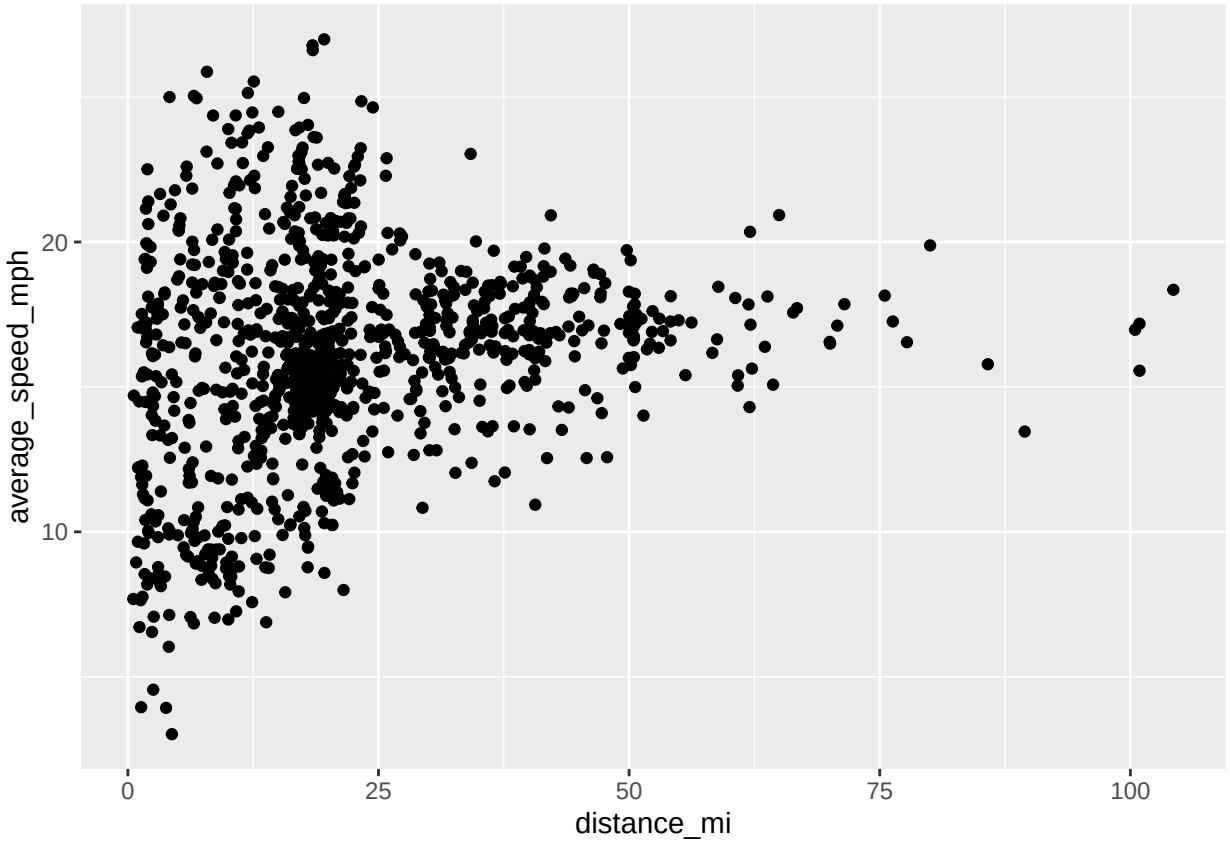



```
##    vars   n   mean    sd median trimmed   mad  min   max range  skew
## X1    1 582 144.18 16.75  146.5  145.28 15.72 77.1 176.6  99.5 -0.64
##    kurtosis   se
## X1     0.19 0.69
## [1] "Kurtosis is 0.19 . Since it's greater than zero, there may be a heavily-tailed distribution. Id
## [1] "Skew is -0.64 . Since it's less than zero, there may be  a pile up of scores on the right of th
```

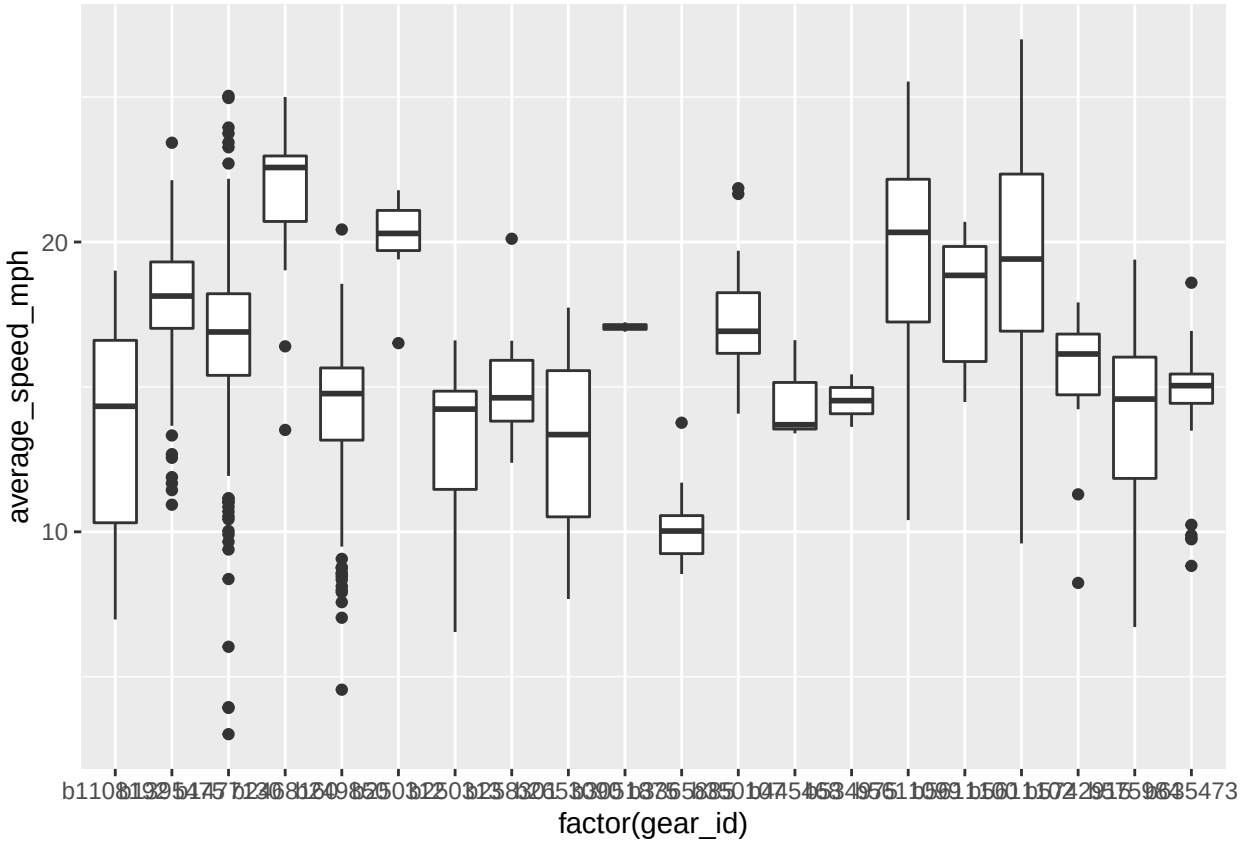
More Analytics

The next couple steps outputs a scatter plot and box plot for distance and spee along with the gear (i.e., bike) used.

```
ggplot(ride_data, aes(x = distance_mi, y = average_speed_mph), color = factor(gear_id)) +
      geom_point()
```




```
# boxplot with bikes
ggplot(data = ride_data, aes(x = factor(gear_id), y = average_speed_mph)) +
        geom_boxplot()
```



Linear Model

A simple linear model is created to show the relationship between average speed and distance. Intuitively, speed should (on average) god down as distance goes up.

```
# Simple lm model - how distance affects speed
lm_speed_dist_gear_id <- lm(average_speed_mph ~ distance_mi + factor(gear_id),
                            data = ride_data)
summary(lm_speed_dist_gear_id)
```

```
##
## Call:
## lm(formula = average_speed_mph ~ distance_mi + factor(gear_id),
##     data = ride_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.8125  -1.4624   0.2481   1.7119   9.1170
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              12.857490   0.326437  39.387  < 2e-16 ***
## distance_mi               0.043389   0.006276   6.913 8.22e-12 ***
## factor(gear_id)b1395475   3.591016   0.395129   9.088  < 2e-16 ***
## factor(gear_id)b1477130   2.782208   0.344437   8.078 1.80e-15 ***
## factor(gear_id)b2468160   7.929199   0.652007  12.161  < 2e-16 ***
## factor(gear_id)b249850    0.279626   0.394636   0.709 0.478750
```

```
## factor(gear_id)b250312   6.597982   1.112628   5.930 4.10e-09 ***
## factor(gear_id)b250313  -0.507563   0.515484  -0.985 0.325031
## factor(gear_id)b258301   0.662742   0.909835   0.728 0.466518
## factor(gear_id)b2653090 -0.674663   0.522756  -1.291 0.197130
## factor(gear_id)b3051875  3.722641   2.025300   1.838 0.066334 .
## factor(gear_id)b3365885 -2.965133   0.819184  -3.620 0.000309 ***
## factor(gear_id)b350107   3.297061   0.720521   4.576 5.31e-06 ***
## factor(gear_id)b445468   0.915452   1.662251   0.551 0.581936
## factor(gear_id)b534975   0.806894   2.024585   0.399 0.690307
## factor(gear_id)b5611099  6.306914   0.676402   9.324  < 2e-16 ***
## factor(gear_id)b5611100  4.549099   1.112942   4.087 4.69e-05 ***
## factor(gear_id)b5611102  6.106592   0.420856  14.510  < 2e-16 ***
## factor(gear_id)b5742915  1.366500   0.814861   1.677 0.093845 .
## factor(gear_id)b575984  -0.017817   0.518860  -0.034 0.972614
## factor(gear_id)b635473   1.104870   0.459090   2.407 0.016271 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.831 on 1051 degrees of freedom
## Multiple R-squared:  0.417,  Adjusted R-squared:  0.4059
## F-statistic: 37.58 on 20 and 1051 DF,  p-value: < 2.2e-16
```

```
lm_speed_dist_gear_id
```

```
##
## Call:
## lm(formula = average_speed_mph ~ distance_mi + factor(gear_id),
##     data = ride_data)
##
## Coefficients:
##             (Intercept)              distance_mi  factor(gear_id)b1395475
##                12.85749                  0.04339                  3.59102
## factor(gear_id)b1477130  factor(gear_id)b2468160   factor(gear_id)b249850
##                 2.78221                  7.92920                  0.27963
##  factor(gear_id)b250312   factor(gear_id)b250313   factor(gear_id)b258301
##                 6.59798                 -0.50756                  0.66274
## factor(gear_id)b2653090  factor(gear_id)b3051875  factor(gear_id)b3365885
##                -0.67466                  3.72264                 -2.96513
##  factor(gear_id)b350107   factor(gear_id)b445468   factor(gear_id)b534975
##                 3.29706                  0.91545                  0.80689
## factor(gear_id)b5611099  factor(gear_id)b5611100  factor(gear_id)b5611102
##                 6.30691                  4.54910                  6.10659
## factor(gear_id)b5742915   factor(gear_id)b575984   factor(gear_id)b635473
##                 1.36650                 -0.01782                  1.10487
```

The linear model shows that distance does not impact speed as expected; it goes up. The lineral model also shows that the gear has a greater impact (depending on the gear), both positively and negatively. This makes sense because a time trial bike will almost always increase speed, whereas a fat tire mountain bike generally slow speed.

Note that there are other factors - e.g., type of terrain - that are not entirely adequately accounted for. Power is probably a better measure overall.

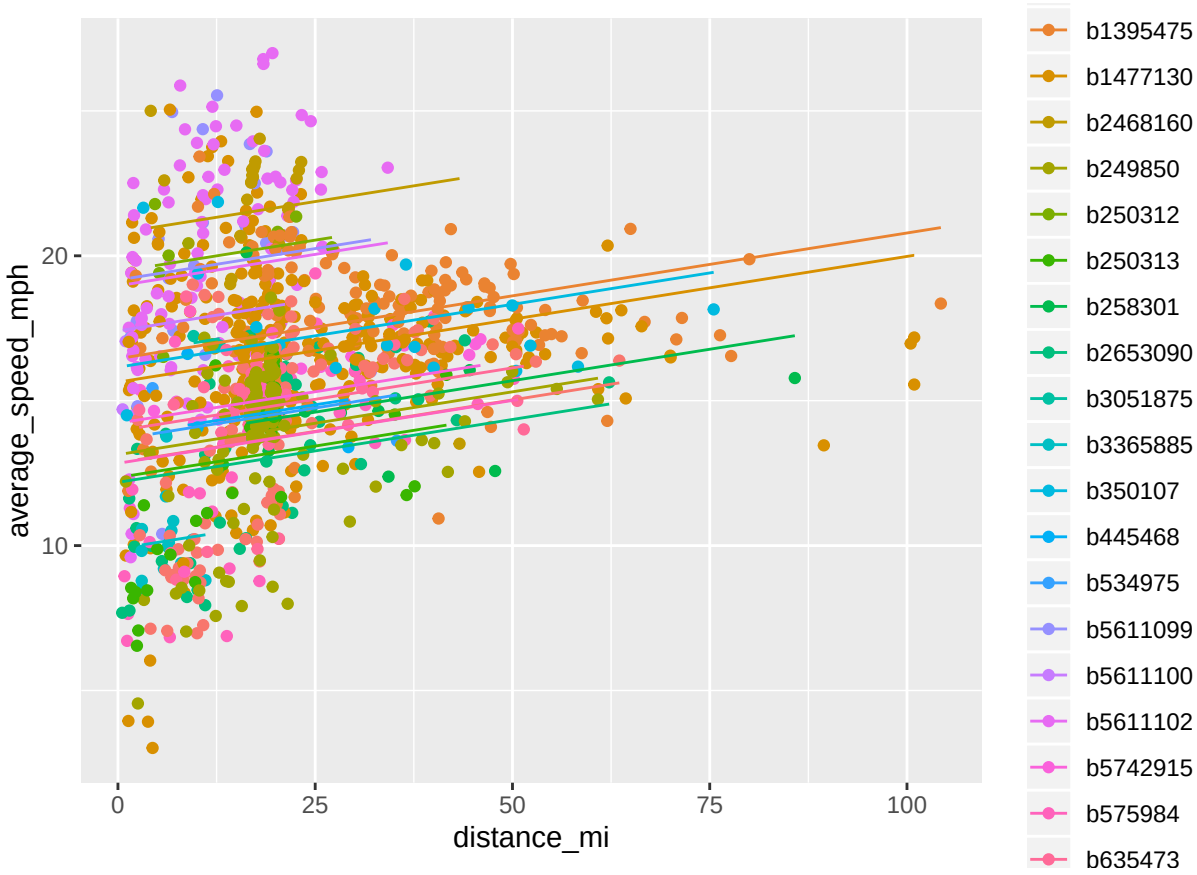
Parallel Slopes

Here we show how the categorical variable (gear) impacts the distance to speed linear model.

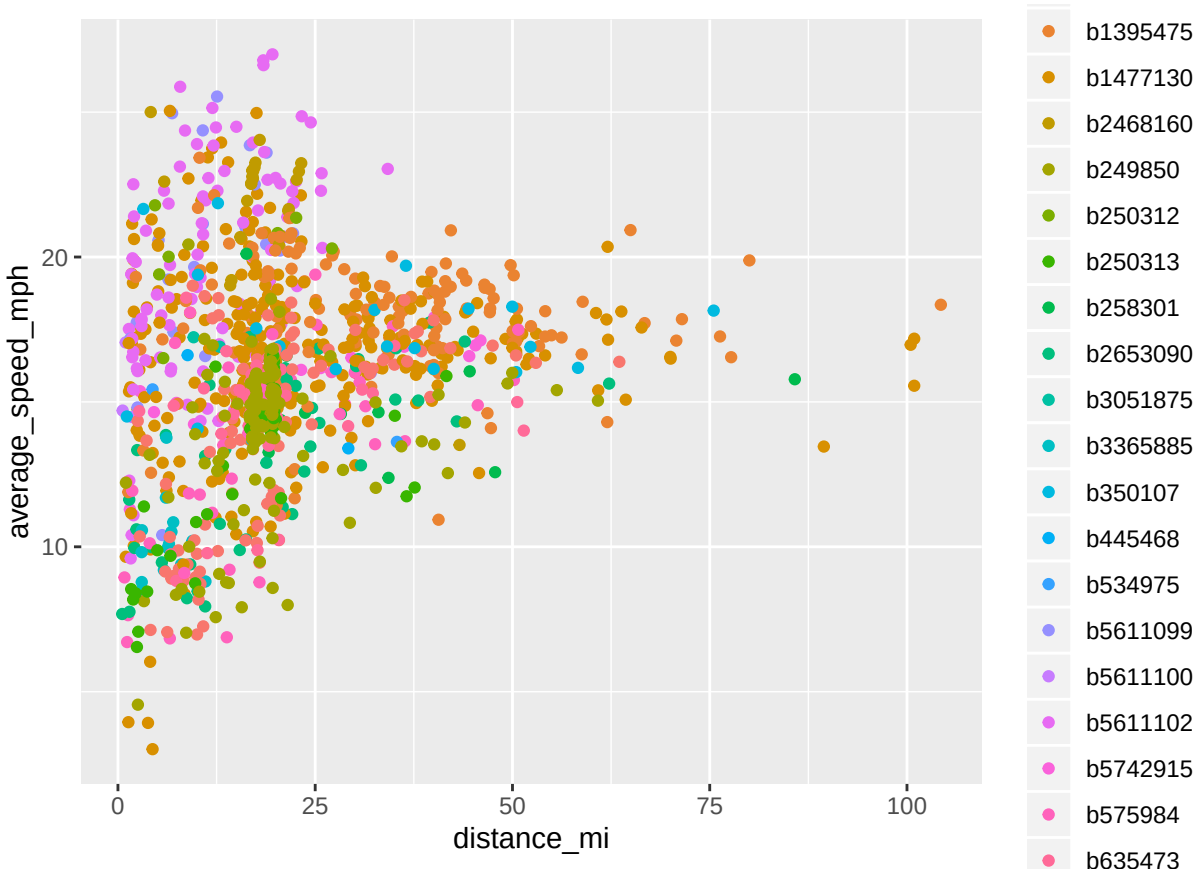
```
# try parallel slopes
# Augment the model
augmented_bikes <- augment(lm_speed_dist_gear_id)

# scatterplot, with color
lm_plot <- ggplot(augmented_bikes, aes(x = distance_mi, y = average_speed_mph,
                                       color = factor.gear_id.)) +
 geom_point()

# single call to geom_line()
lm_plot + geom_line(aes(y = .fitted))
```




```
print(lm_plot)
```


average_speed_mph
20
10
0
25
50
75
100
distance_mi
b1395475
b1477130
b2468160
b249850
b250312
b250313
b258301
b2653090
b3051875
b3365885
b350107
b445468
b534975
b5611099
b5611100
b5611102
b5742915
b575984
b635473