How Much Did It Rain? Part I

Tyler Byers

November 16, 2015

About

This is Part 1 of the final project for the University of Washington Professional and Continuing Education Data Science certificate class #3 of 3. I chose the Kaggle competition "How Much Did it Rain" (https://www.kaggle.com/c/how-much-did-it-rain-ii). This deliverable is mostly telling about the project and exploring the data a little bit. I am working on this project alone – no team members.

Problem Description and Summary

This competition is to predict how much it rained in a given location based on radar readings. Because the amount of rainfall can be highly localized to an area, and because it is impossible to put a rain gauge in every location, or even in a lot of different locations, calculating the amount of rainfall a certain location received can be important. Since these data in particular are for midwestern corn-growing states, estimating the amount of rain a certain farm received can be important – whether to check for crop damage, or to plan harvest, or perhaps do localized flooding mitigation. If we can use radar to accurately calculate the rainfall in a given location, it may help the farmers with these problems. This has applicability to more than just farms too.

This is a supervised learning problem, because we have a target variable `Expected` amount of rain in mm in an hour. We use the radar measurements to predict the target variable. The evaluation metric is Mean Absolute Error (MAE) for the Kaggle competition.

As described below, the main thing that I'll have to do with the data when doing my prediction is to filter out bad values for `Expected` – probably those values with readings above 350 mm in an hour. As described on Kaggle, the gauges can often give faulty readings. There will be a bit of work involved with all the NA data – not sure what I'll do with that yet. Choosing the correct algorithms will be non-trivial, and I'll have to really think hard about how to aggregate the data since there may be anywhere between one and nineteen radar observations in an hour.

Set Environment

First we need to set our environment to the proper working directory and load the needed R packages.

```
##-----Set working directory-----
setwd('~/UW_DataScience/DataAtScale/final_project/')

##-----Load Libraries-----
library(dplyr); library(ggplot2); library(readr)
```

Load Data, Perform EDA

Now we'll load the data and explore it a bit.

```
#rain <- read.csv('./data/train.csv')  # was throwing an error when knitting/caching PDF
rain <- readRDS('./data/rain.rds') # converted file to RDS to load here
```

```
dim(rain)
```

```
## [1] 13765201       24
```

```
str(rain)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    13765201 obs. of  24 variables:
##  $ Id                    : int  1 1 1 1 1 1 2 2 2 2 ...
##  $ minutes_past          : int  3 16 25 35 45 55 1 6 11 16 ...
##  $ radardist_km          : num  10 10 10 10 10 10 2 2 2 2 ...
##  $ Ref                   : num  NA NA NA NA NA NA 9 26.5 21.5 18 ...
##  $ Ref_5x5_10th          : num  NA NA NA NA NA NA 5 22.5 15.5 14 ...
##  $ Ref_5x5_50th          : num  NA NA NA NA NA NA 7.5 25.5 20.5 17.5 ...
##  $ Ref_5x5_90th          : num  NA NA NA NA NA NA 10.5 31.5 25 21 ...
##  $ RefComposite          : num  NA NA NA NA NA NA 15 26.5 26.5 20.5 ...
##  $ RefComposite_5x5_10th: num  NA NA NA NA NA NA 10.5 26.5 23.5 18 ...
##  $ RefComposite_5x5_50th: num  NA NA NA NA NA NA 16.5 28.5 25 20.5 ...
##  $ RefComposite_5x5_90th: num  NA NA NA NA NA NA 23.5 32 27 23 ...
##  $ RhoHV                 : num  NA NA NA NA NA ...
##  $ RhoHV_5x5_10th        : num  NA NA NA NA NA ...
##  $ RhoHV_5x5_50th        : num  NA NA NA NA NA ...
##  $ RhoHV_5x5_90th        : num  NA NA NA NA NA ...
##  $ Zdr                   : num  NA NA NA NA NA ...
##  $ Zdr_5x5_10th          : num  NA NA NA NA NA ...
##  $ Zdr_5x5_50th          : num  NA NA NA NA NA ...
##  $ Zdr_5x5_90th          : num  NA NA NA NA NA ...
##  $ Kdp                   : num  NA NA NA NA NA ...
##  $ Kdp_5x5_10th          : num  NA NA NA NA NA ...
##  $ Kdp_5x5_50th          : num  NA NA NA NA NA ...
##  $ Kdp_5x5_90th          : num  NA NA NA NA NA ...
##  $ Expected              : num  0.254 0.254 0.254 0.254 0.254 ...
```

```
summary(rain)
```

```
##        Id           minutes_past    radardist_km        Ref
##  Min.   :      1   Min.   : 0.00   Min.   : 0.00   Min.   :-31
##  1st Qu.: 296897   1st Qu.:15.00   1st Qu.: 9.00   1st Qu.: 16
##  Median : 592199   Median :30.00   Median :11.00   Median : 22
##  Mean   : 592337   Mean   :29.52   Mean   :11.07   Mean   : 23
##  3rd Qu.: 889582   3rd Qu.:44.00   3rd Qu.:14.00   3rd Qu.: 30
##  Max.   :1180945   Max.   :59.00   Max.   :21.00   Max.   : 71
##                                                    NA's   :7415826
##   Ref_5x5_10th      Ref_5x5_50th      Ref_5x5_90th      RefComposite
##  Min.   :-32       Min.   :-32       Min.   :-28       Min.   :-32
##  1st Qu.: 14       1st Qu.: 16       1st Qu.: 18       1st Qu.: 18
##  Median : 20       Median : 22       Median : 26       Median : 24
##  Mean   : 20       Mean   : 23       Mean   : 26       Mean   : 25
##  3rd Qu.: 26       3rd Qu.: 29       3rd Qu.: 34       3rd Qu.: 32
##  Max.   : 62       Max.   : 69       Max.   : 72       Max.   : 92
##  NA's   :8481213   NA's   :7408719   NA's   :6213920   NA's   :7048858
##  RefComposite_5x5_10th RefComposite_5x5_50th RefComposite_5x5_90th
##  Min.   :-31           Min.   :-28           Min.   :-25
```

```
##  1st Qu.: 16          1st Qu.: 18          1st Qu.: 20
##  Median : 22          Median : 24          Median : 27
##  Mean   : 22          Mean   : 24          Mean   : 27
##  3rd Qu.: 28          3rd Qu.: 32          3rd Qu.: 35
##  Max.   : 66          Max.   : 71          Max.   : 94
##  NA's   :8009528      NA's   :7053538      NA's   :5935998
##      RhoHV         RhoHV_5x5_10th    RhoHV_5x5_50th    RhoHV_5x5_90th
##  Min.   :0         Min.   :0         Min.   :0         Min.   :0
##  1st Qu.:1         1st Qu.:1         1st Qu.:1         1st Qu.:1
##  Median :1         Median :1         Median :1         Median :1
##  Mean   :1         Mean   :1         Mean   :1         Mean   :1
##  3rd Qu.:1         3rd Qu.:1         3rd Qu.:1         3rd Qu.:1
##  Max.   :1         Max.   :1         Max.   :1         Max.   :1
##  NA's   :8830285   NA's   :9632047   NA's   :8828633   NA's   :7859617
##       Zdr           Zdr_5x5_10th      Zdr_5x5_50th      Zdr_5x5_90th
##  Min.   :-8        Min.   :-8        Min.   :-8        Min.   :-8
##  1st Qu.: 0        1st Qu.:-1        1st Qu.: 0        1st Qu.: 1
##  Median : 0        Median :-1        Median : 0        Median : 2
##  Mean   : 1        Mean   :-1        Mean   : 0        Mean   : 2
##  3rd Qu.: 1        3rd Qu.: 0        3rd Qu.: 1        3rd Qu.: 3
##  Max.   : 8        Max.   : 8        Max.   : 8        Max.   : 8
##  NA's   :8830285   NA's   :9632047   NA's   :8828633   NA's   :7859617
##       Kdp           Kdp_5x5_10th       Kdp_5x5_50th      Kdp_5x5_90th
##  Min.   :-96       Min.   :-81        Min.   :-79       Min.   :-100
##  1st Qu.: -1       1st Qu.: -5        1st Qu.: -1       1st Qu.:   2
##  Median :  0       Median : -3        Median :  0       Median :   4
##  Mean   :  0       Mean   : -3        Mean   :  0       Mean   :   4
##  3rd Qu.:  2       3rd Qu.: -2        3rd Qu.:  0       3rd Qu.:   6
##  Max.   :180       Max.   :  4        Max.   : 13       Max.   : 145
##  NA's   :9582566   NA's   :10336419   NA's   :9577920   NA's   :8712425
##     Expected
##  Min.   :    0.01
##  1st Qu.:    0.25
##  Median :    1.02
##  Mean   :  108.63
##  3rd Qu.:    3.81
##  Max.   :33017.73
##
```

I notice with the summary that a very high percentage of several of the variables are NA. So we're going to either have to choose a prediction algorithm that can handle NA values, or get rid of those values, or handle them some other way.

Explore Target Variable

One interesting thing about the data is that the target variable, `Expected` is repeated several times for each `Id`. This is because there are several radar observations per hour per `Id` but only one `Expected` reading, which is the millimeters of rain in that observation hour.

So, in order to see what our distribution of rain measurements actually looks like, we have to take just a single value per `Id`. We are going to take the first value as well as the mean Expected value for each Id. If they are not the same, then there is some problem with the data, because all the values per Id chunk should be identical.

```
rain_1perhr <- rain %>%
  group_by(Id) %>%
  summarise(Expected = first(Expected), Expected_avg = mean(Expected), n = n())
```

How many of the `Expected` values were different than the mean `Expected` per `Id` chunk?

```
sum(rain_1perhr$Expected != rain_1perhr$Expected_avg)
```

```
## [1] 0
```

The sum is zero, so they were all the same, so there are no abnormalities with that data.

How many radar observations in a given hour are we seeing? By table:

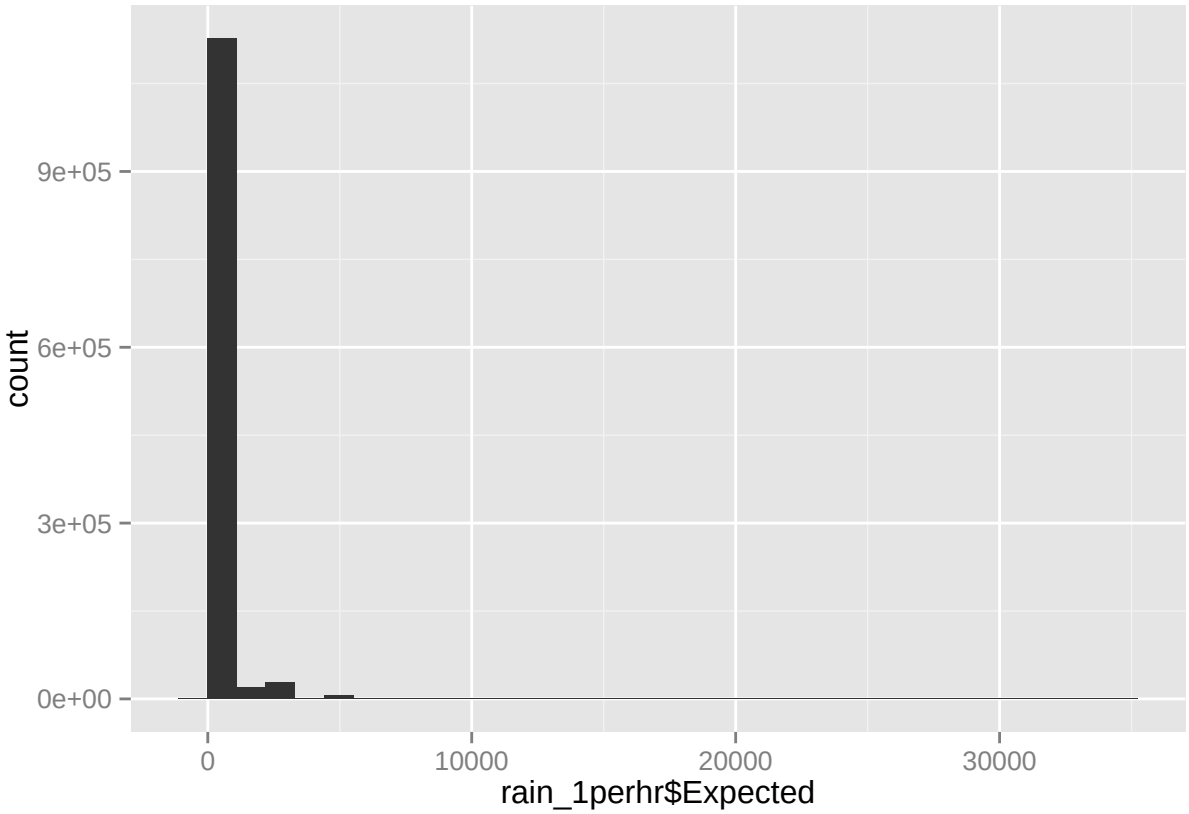
```
table(rain_1perhr$n)
```

```
##
##      1      2      3      4      5      6      7      8      9     10
##    220    354    288    217   1300 152821  50919  10508  16555 134745
##     11     12     13     14     15     16     17     18     19
## 138486 163082 191744 134242  56683  49385  41379  16388  21629
```

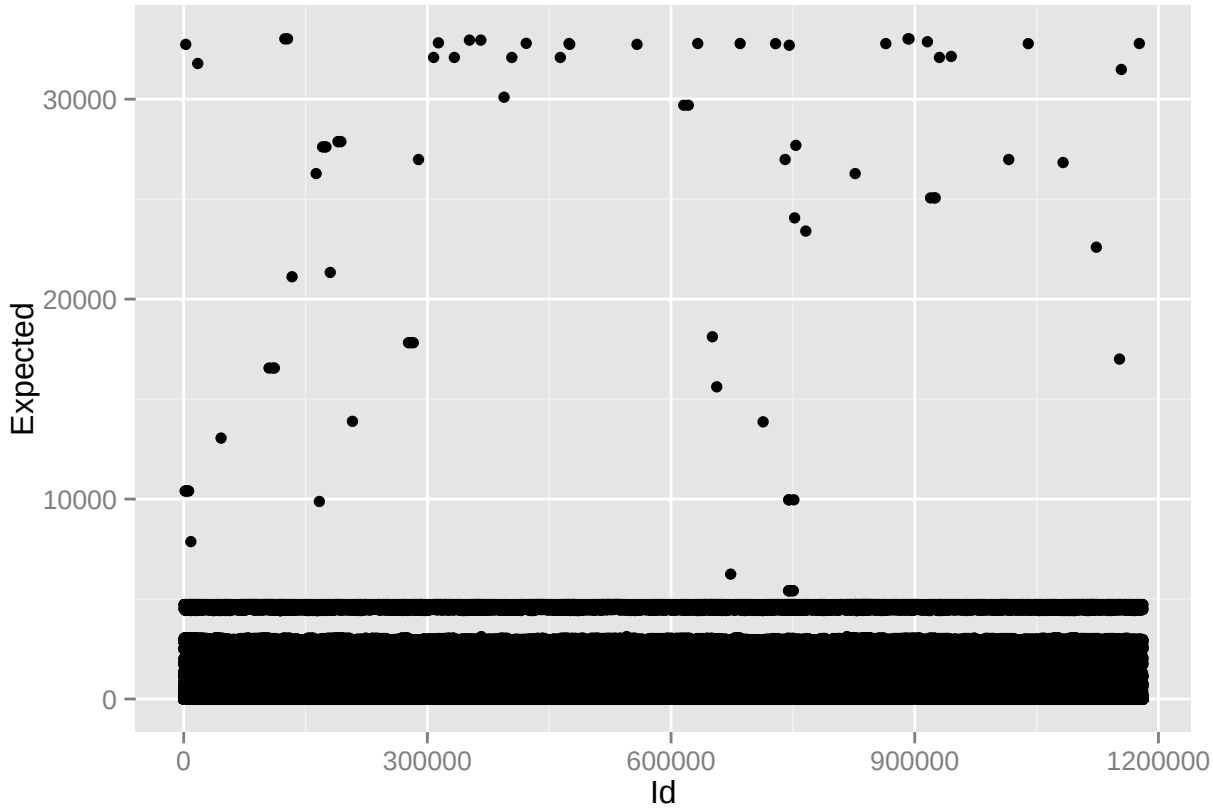
This is interesting – there may be different ways of handling the 1-observation hour versus the 19-observation hours. Maybe by level of certainty/uncertainty. Not sure, will have to think about this.

Now plot the Expected.

```
qplot(rain_1perhr$Expected)
```



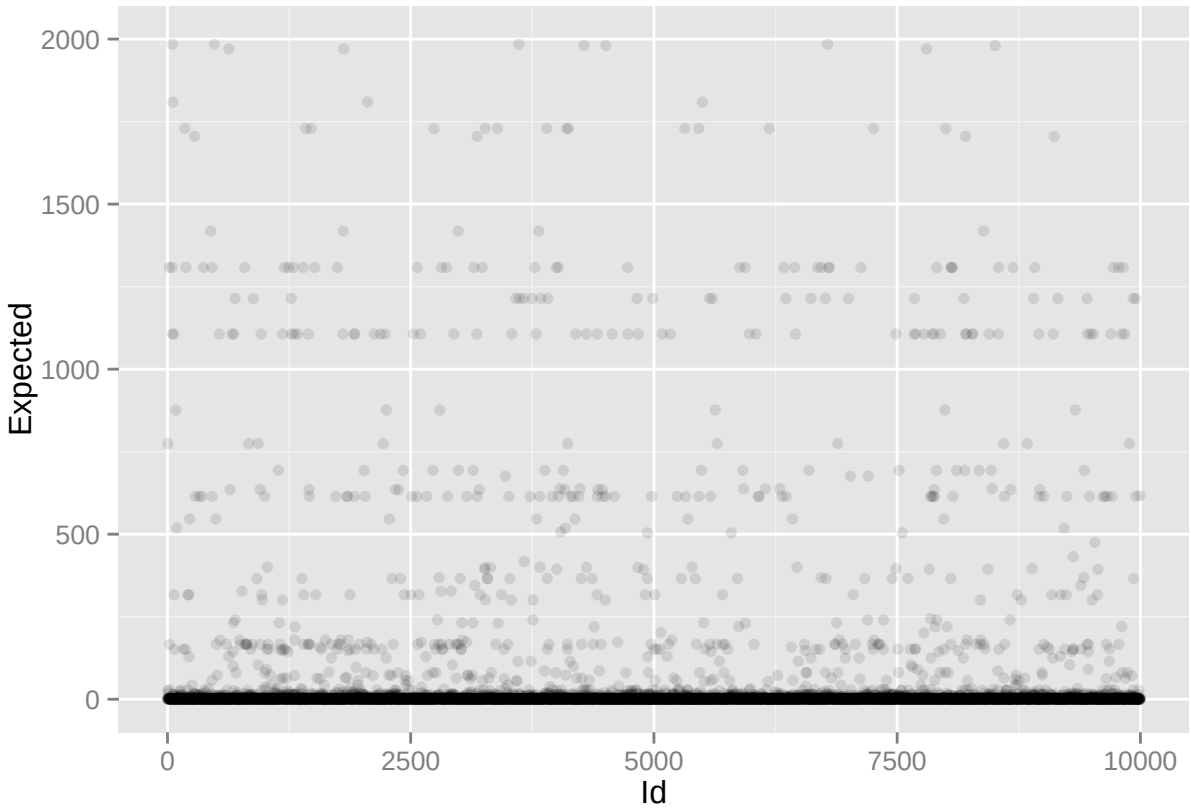
```
ggplot(rain_1perhr, aes(x = Id, y = Expected)) + geom_point()
```



That plot is extremely difficult to see. I susupect there is a lot of bad data in this data set. Let's zoom in a bit:

```
ggplot(rain_1perhr %>% filter(Id < 10000), aes(x = Id, y = Expected)) + geom_point(alpha = 0.1) +
  scale_y_continuous(limits  = c(0, 2000))
```

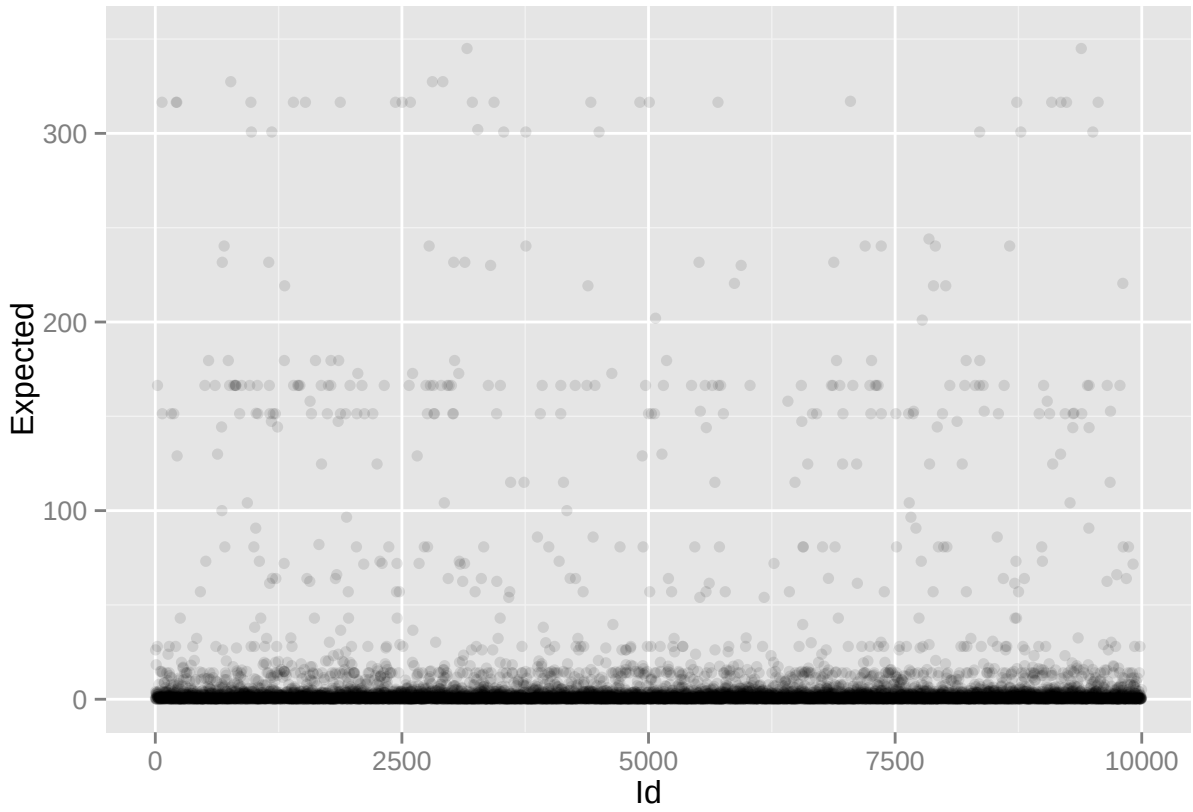
```
## Warning in loop_apply(n, do.ply): Removed 314 rows containing missing
## values (geom_point).
```



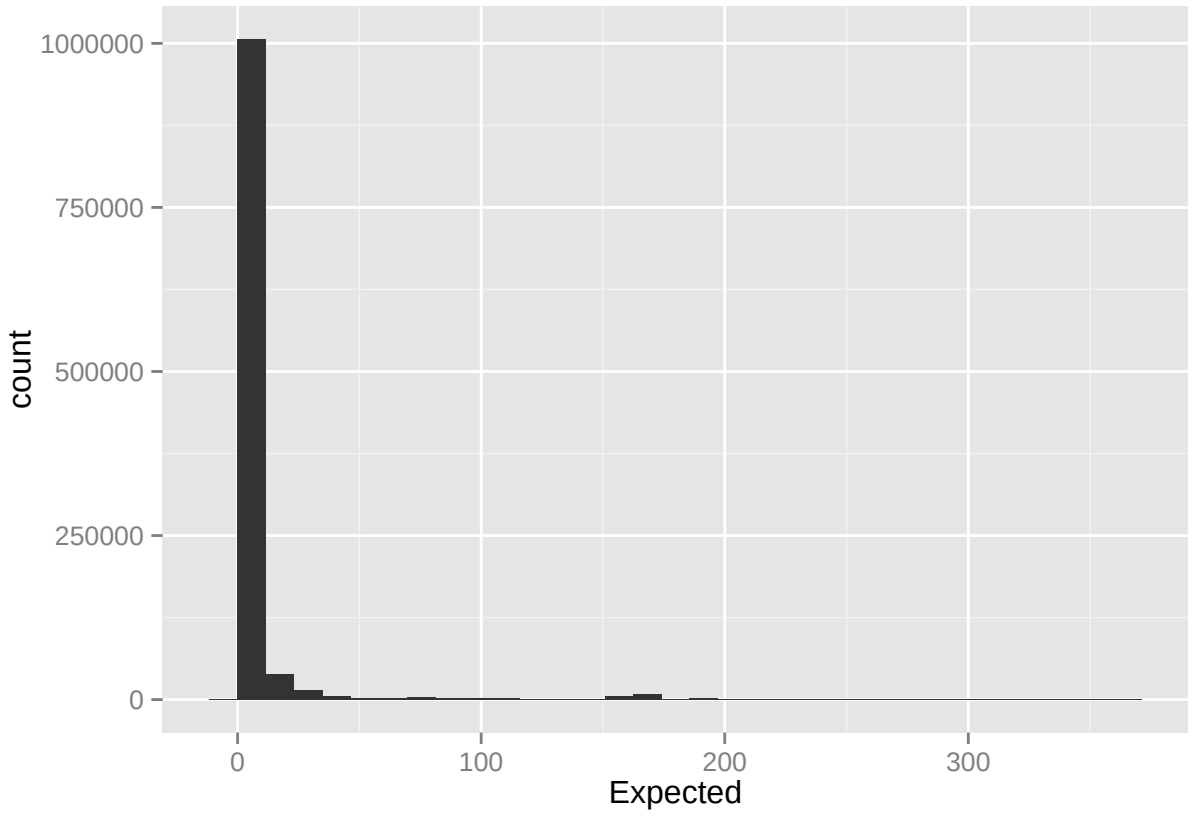
In fact, according to the Weather Channel, the highest 1-hour rainfall total in the USA, ever, was 13.8 inches in an hour (http://www.weather.com/holiday/spring/news/extreme-rainfall-records-united-states-20130313#/4). This equates to 350 mm. So, we can probably throw out rainfall values above 350mm, and consider them to be bad gauge data.

```
ggplot(rain_1perhr %>% filter(Id < 10000), aes(x = Id, y = Expected)) + geom_point(alpha = 0.1) + scale_
```

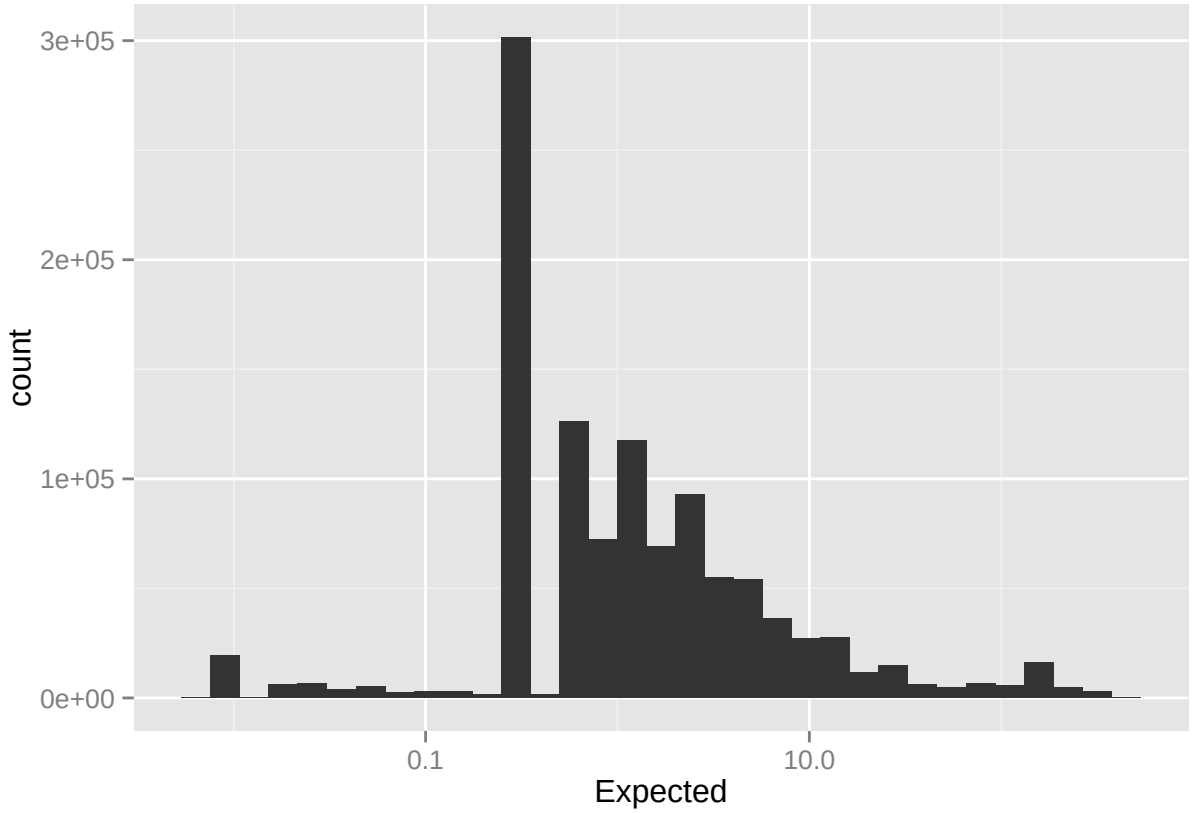
```
## Warning in loop_apply(n, do.ply): Removed 636 rows containing missing
## values (geom_point).
```




```
ggplot(rain_1perhr %>% filter(Expected < 350), aes(x = Expected)) + geom_histogram()
```



```
ggplot(rain_1perhr %>% filter(Expected < 350), aes(x = Expected)) + geom_histogram() +
  scale_x_log10()
```



When filtering to allow no values of `Expected` greater than 350, the scatterplot reveals, unsurprisingly, a skew-right pattern. The second histogram above is a log-10 scale, which is closer to a normal type of distribution. So, one way that we might be more successful at solving this problem is to treat expected as a log and then re-scale it for the final answer.

Just curious, how many values to we "screen out" when we filter to reasonable rain values? Beyond screening out completely ridiculous values, it's hard to screen out bad values that lie within a reasonable range.

```
nrow(rain_1perhr %>% filter(Expected < 350))
```

```
## [1] 1107673
```

```
nrow(rain_1perhr %>% filter(Expected < 350))/nrow(rain_1perhr)
```

```
## [1] 0.9379548
```

So, at this leaves us with about 93.8% of our original data.

Looking at Individual Variables

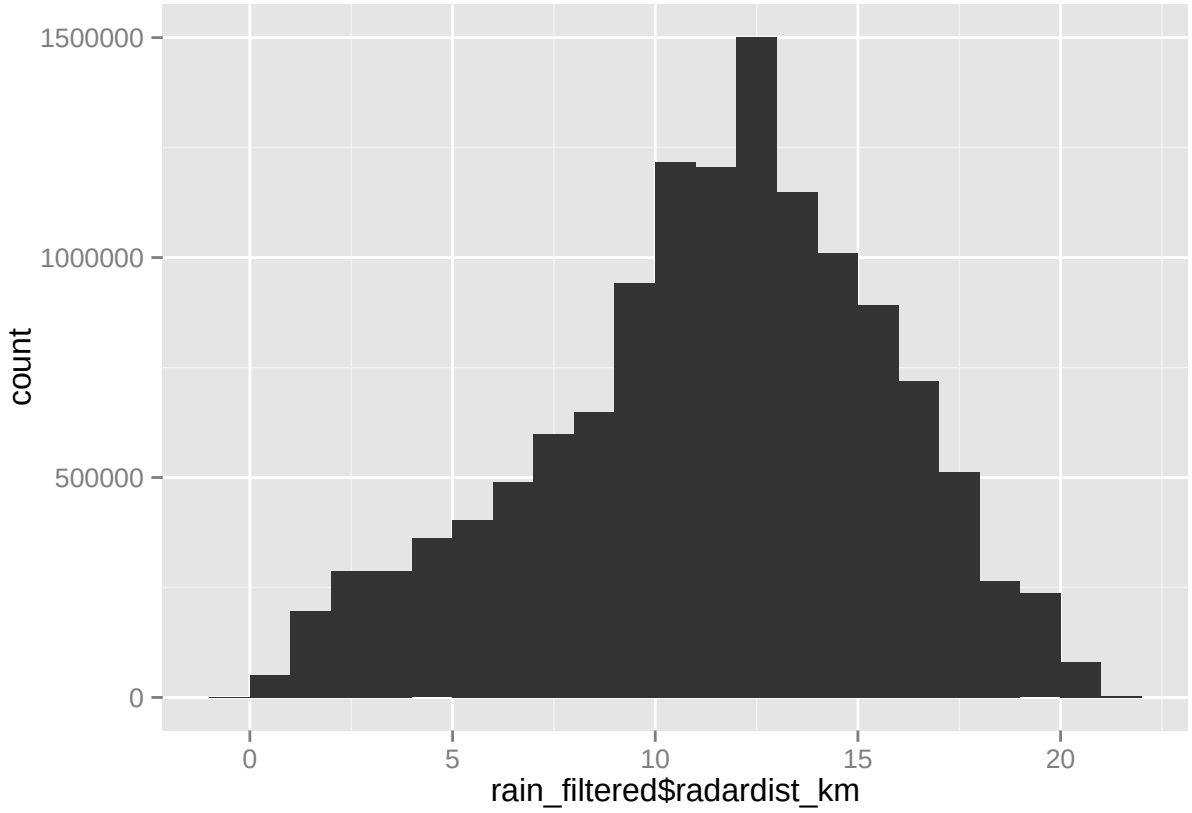
Now I will, tediously, look at histograms for the remaining variables, to see if they look reasonable and maybe if we should think about applying transformations to the data when doing our predictions.

First, need to filter out the "bad" rain gauge values.

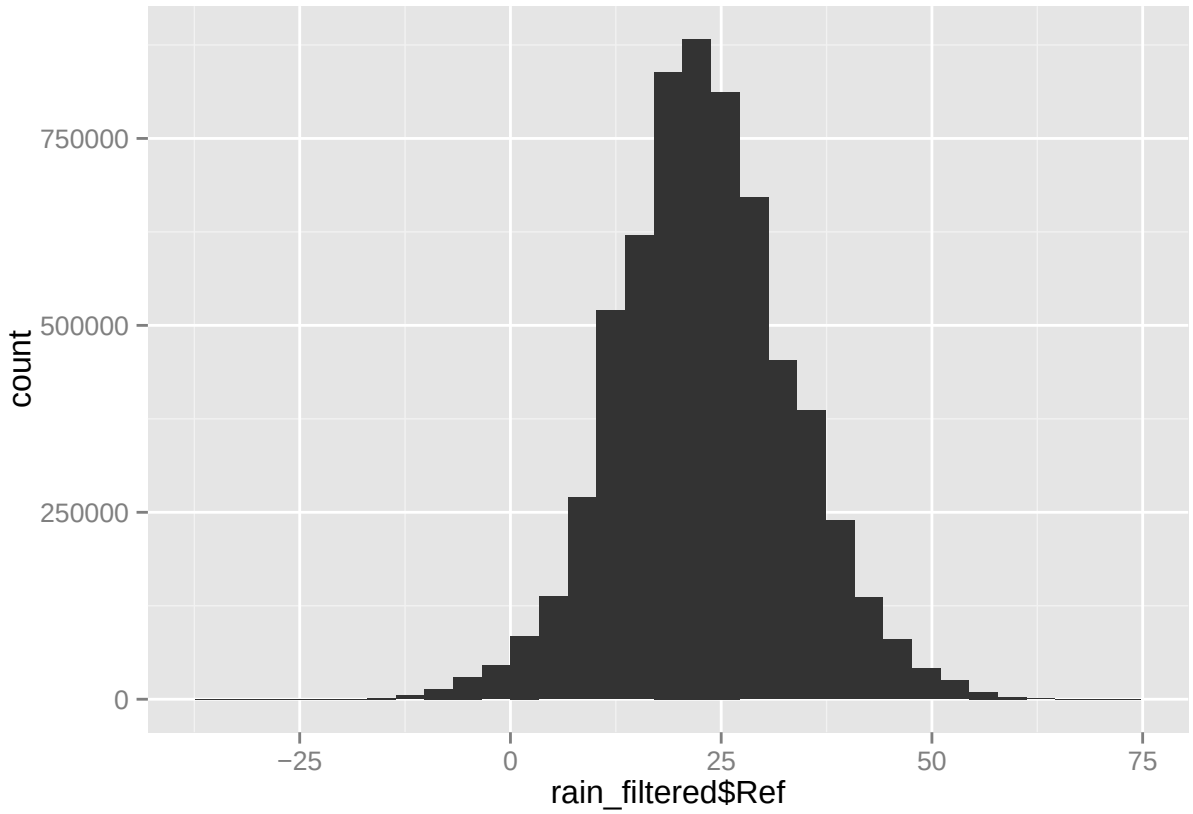
```
rain_filtered <- rain %>% filter(Expected < 350)
```

Now, histograms for each variable remaining. Mainly looking for outliers and distribution shapes. Comments about the distribution shapes are included in-line.

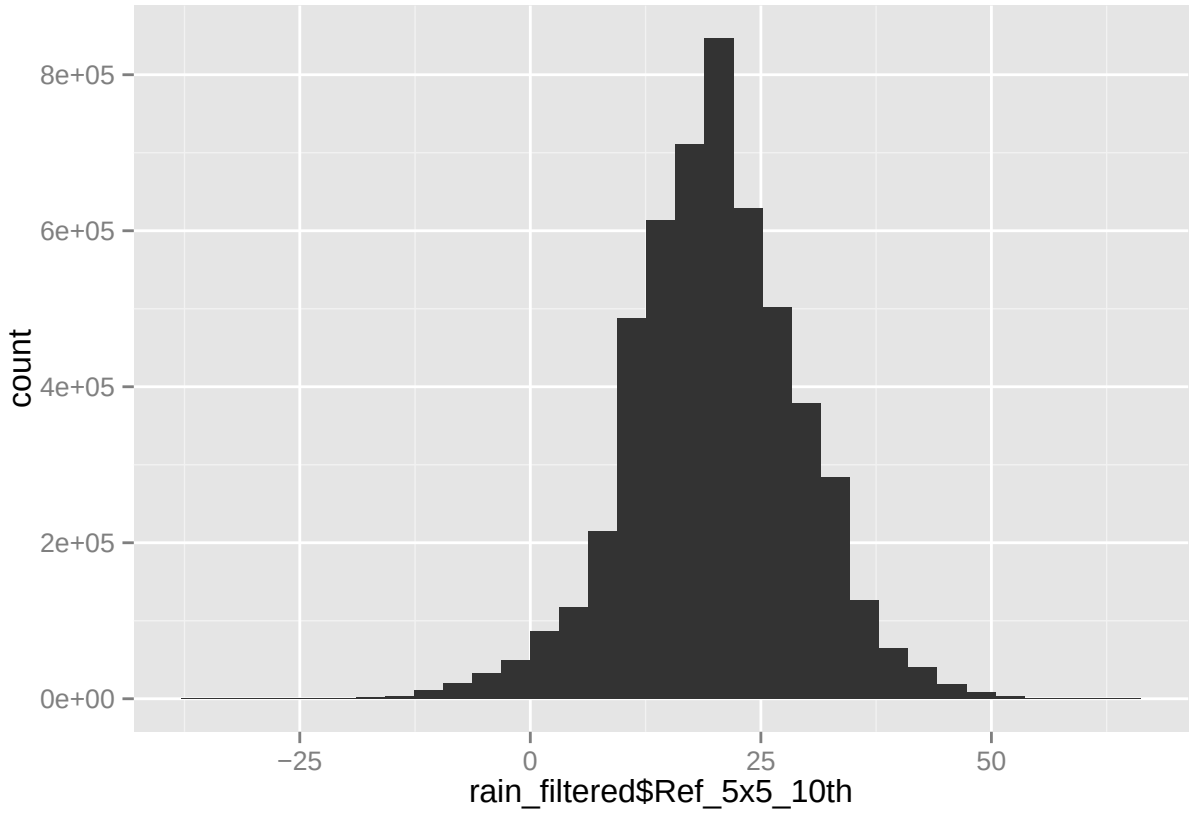
```
qplot(rain_filtered$radardist_km, binwidth = 1)  # normal-ish distribution, little skew left
```



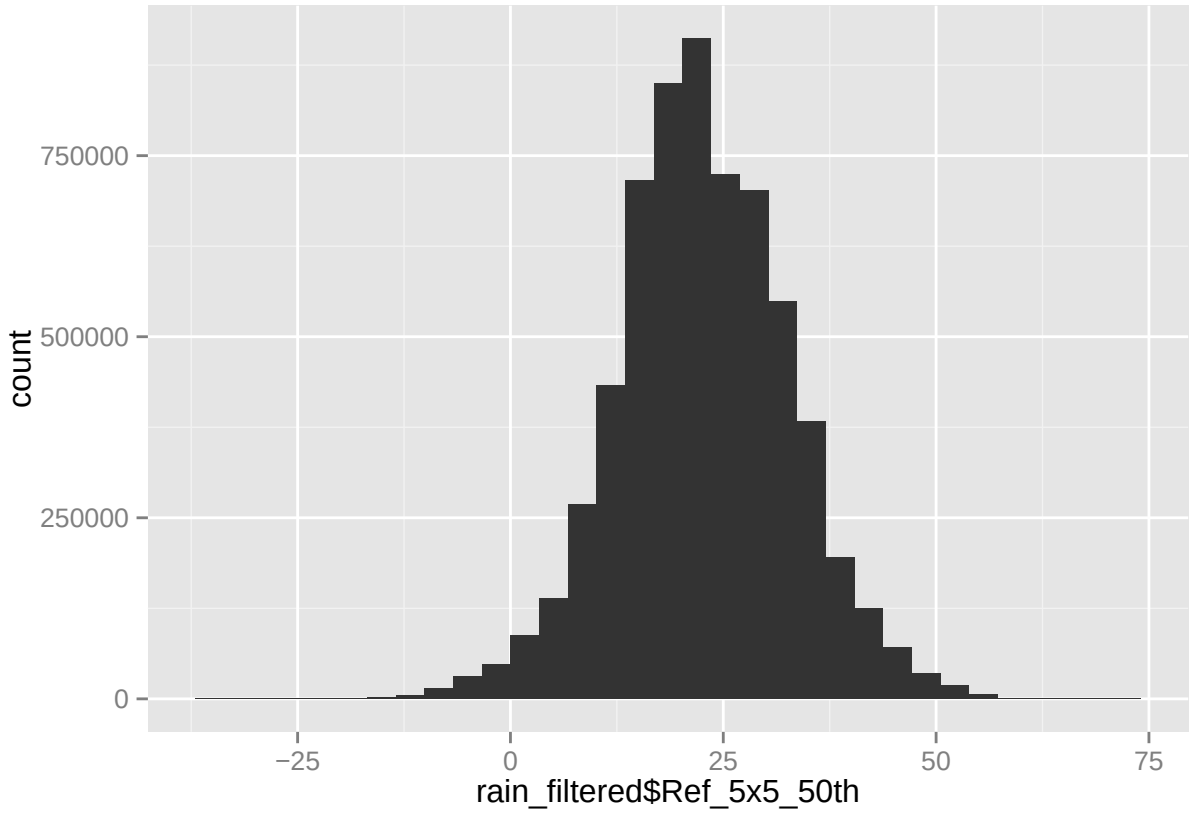

```
qplot(rain_filtered$Ref) # normal dist
```



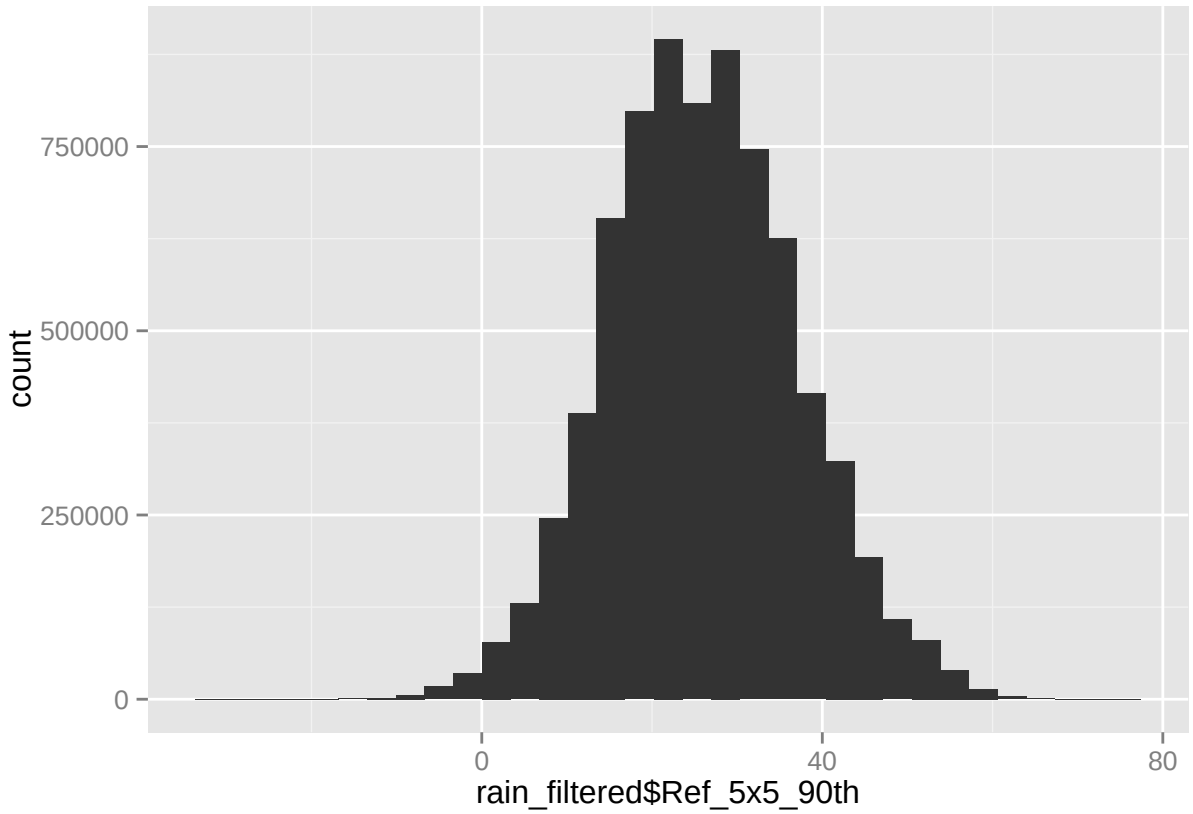

```
qplot(rain_filtered$Ref_5x5_10th) # normal dist
```



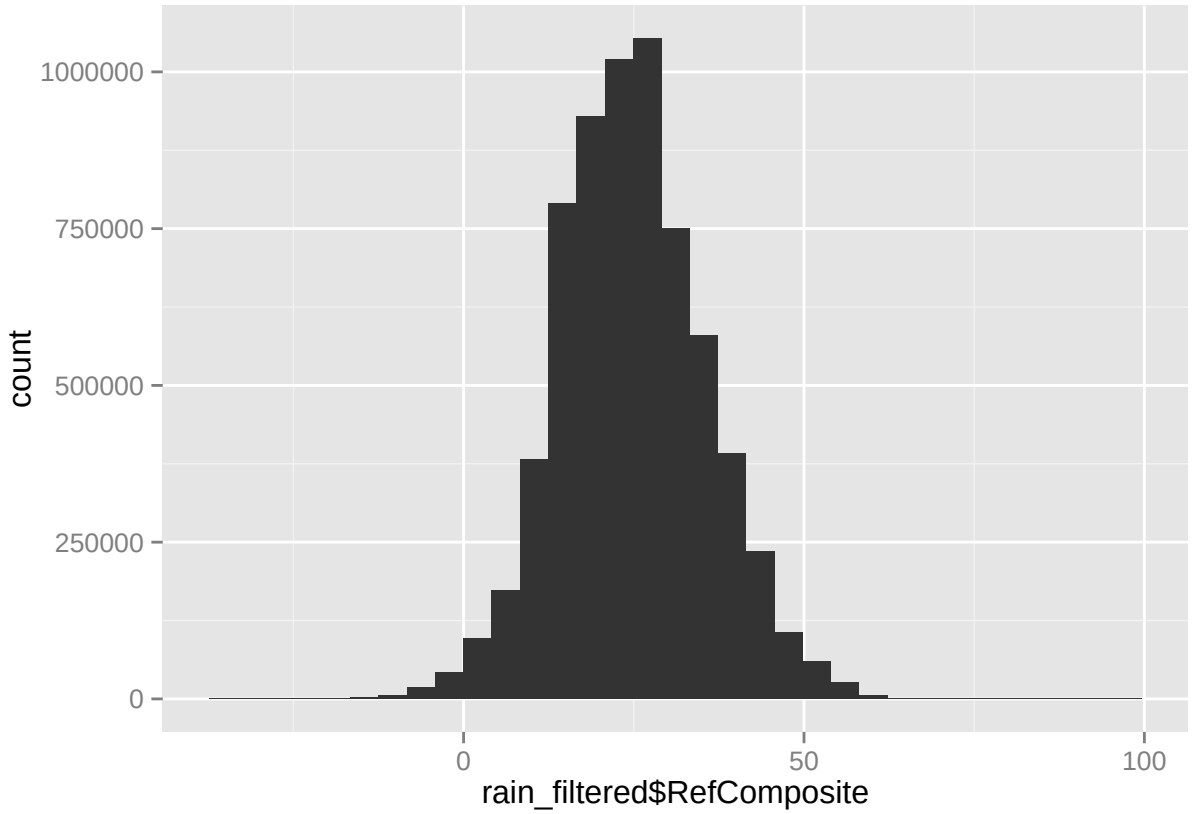
```
qplot(rain_filtered$Ref_5x5_50th) # normal dist
```



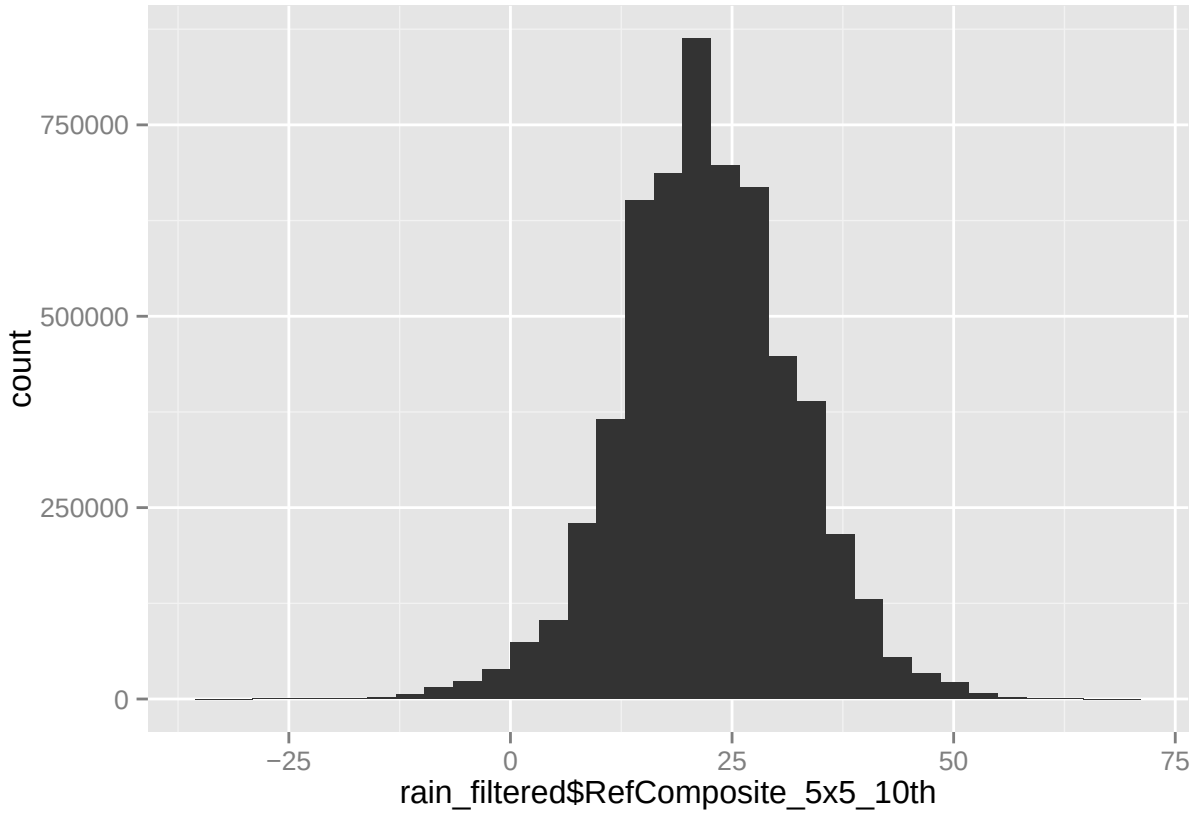

```
qplot(rain_filtered$Ref_5x5_90th) # normal dist
```



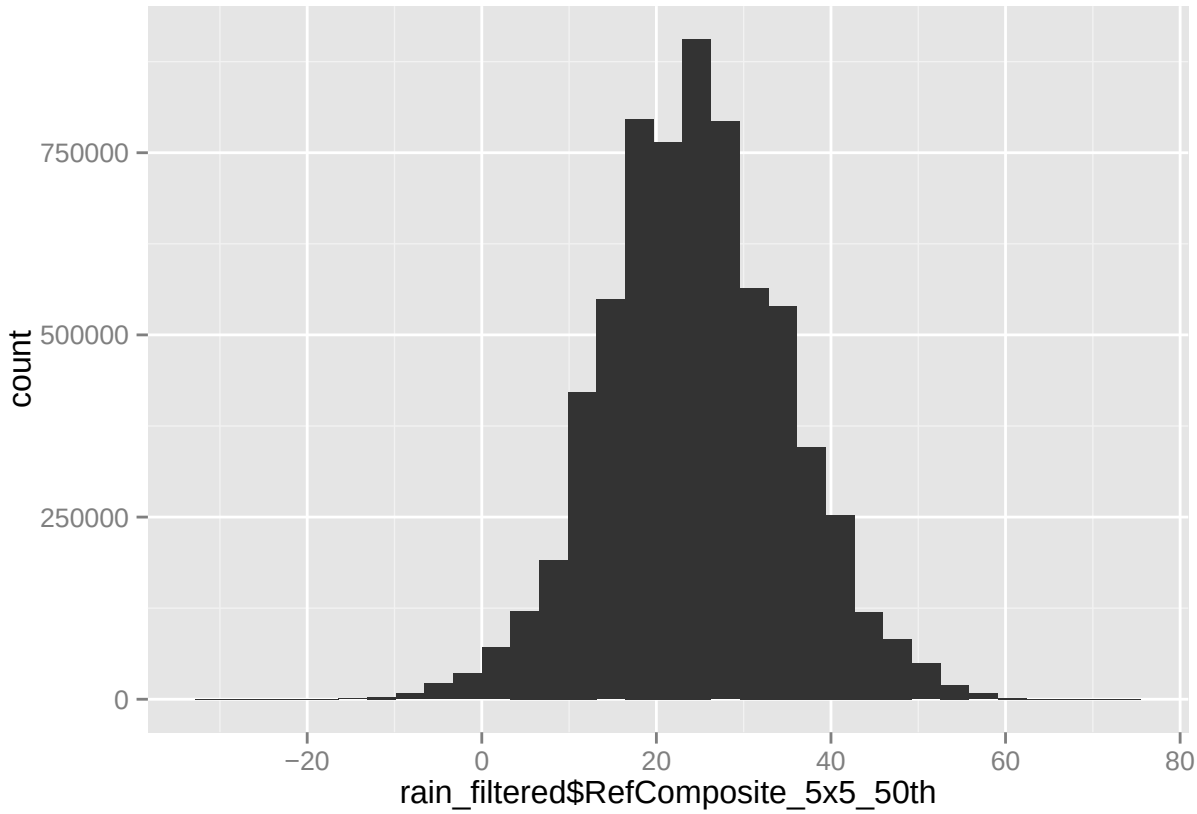

```
qplot(rain_filtered$RefComposite) # normal dist
```



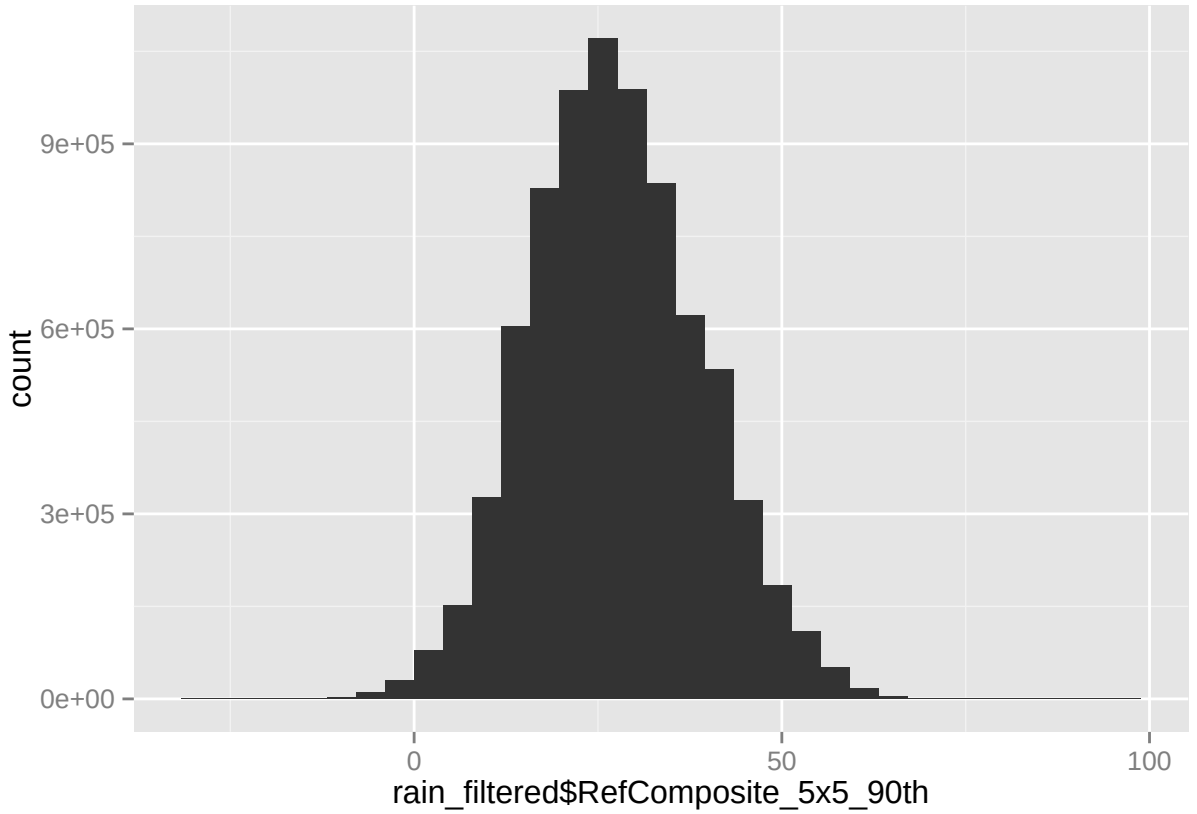
```
qplot(rain_filtered$RefComposite_5x5_10th) # normal dist
```



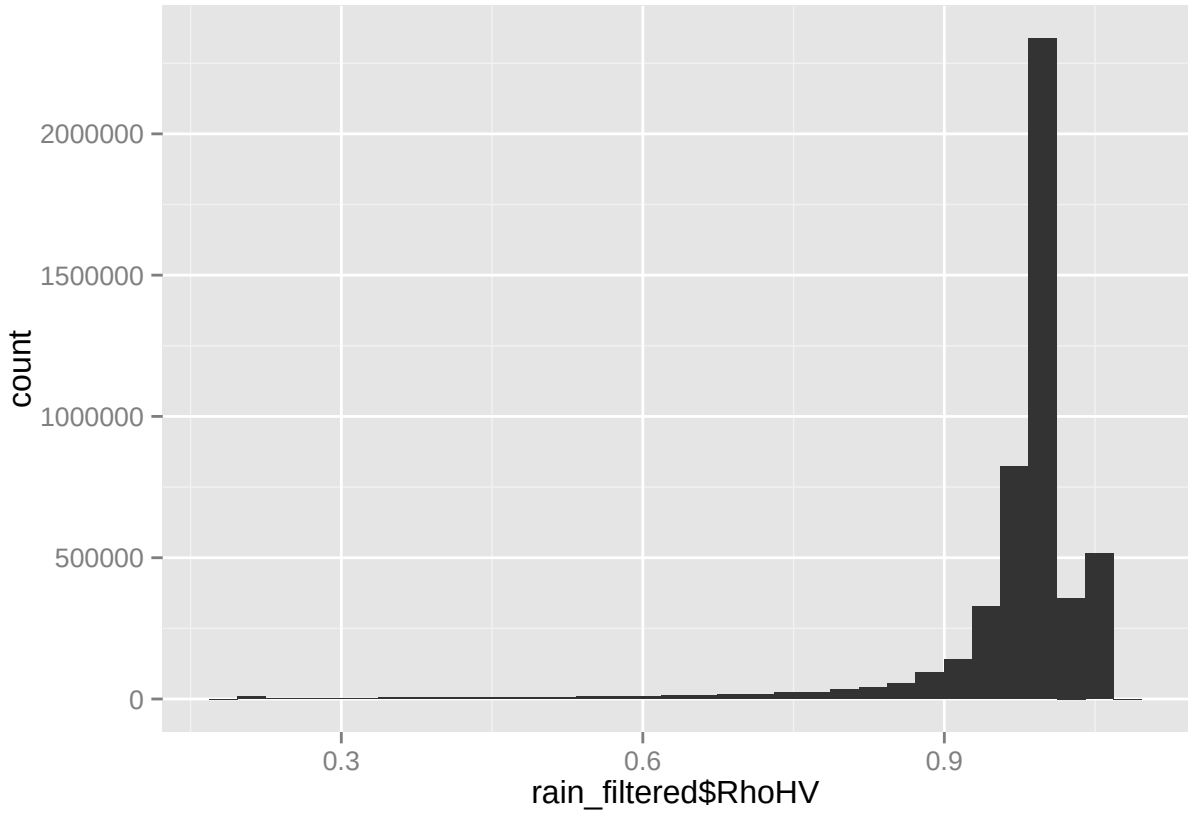

```
qplot(rain_filtered$RefComposite_5x5_50th) # normal dist
```



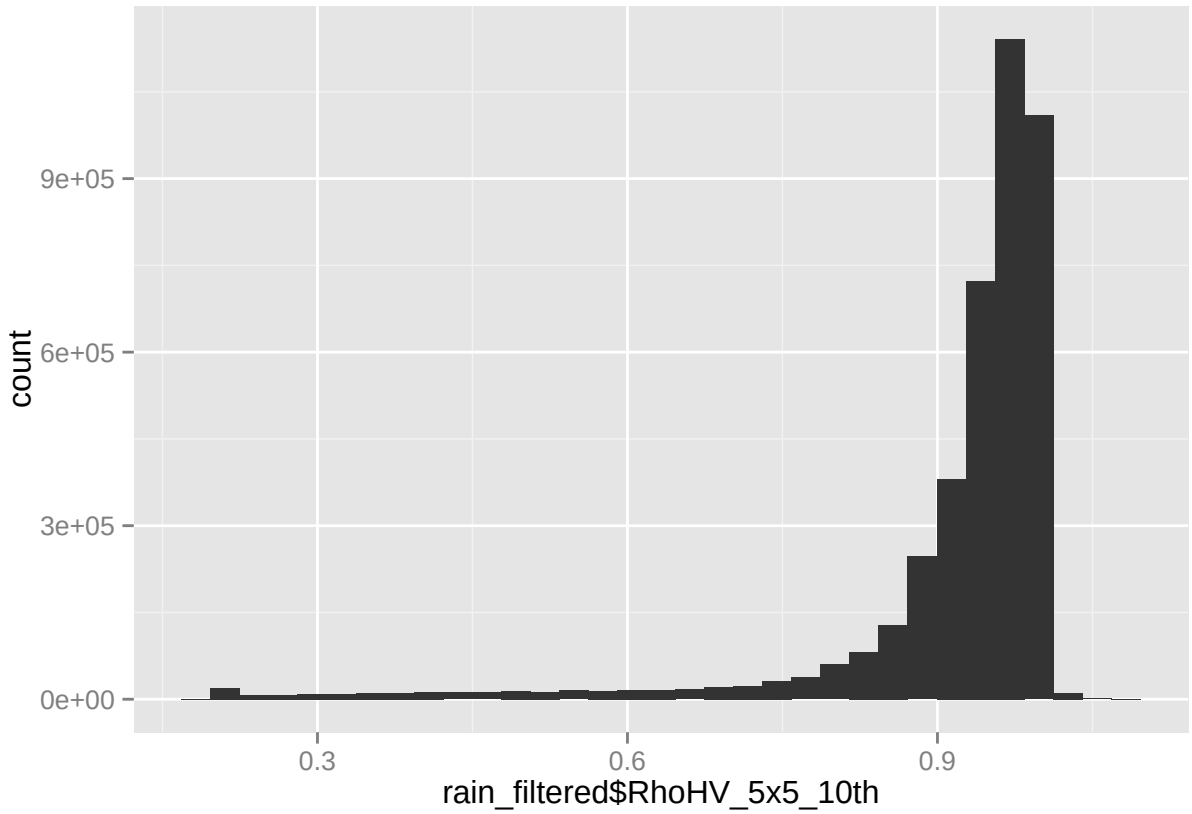

```
qplot(rain_filtered$RefComposite_5x5_90th) # normal dist
```



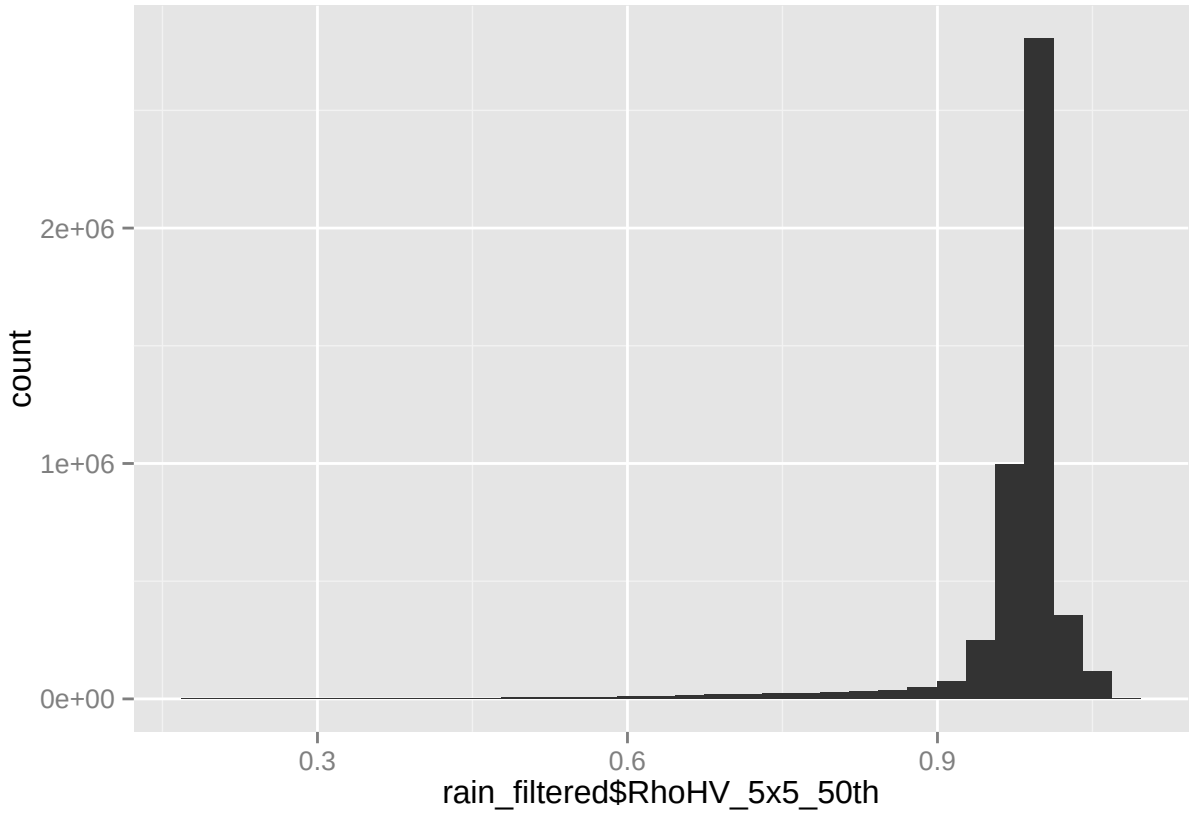
```
qplot(rain_filtered$RhoHV)  # very skew-left distribution
```



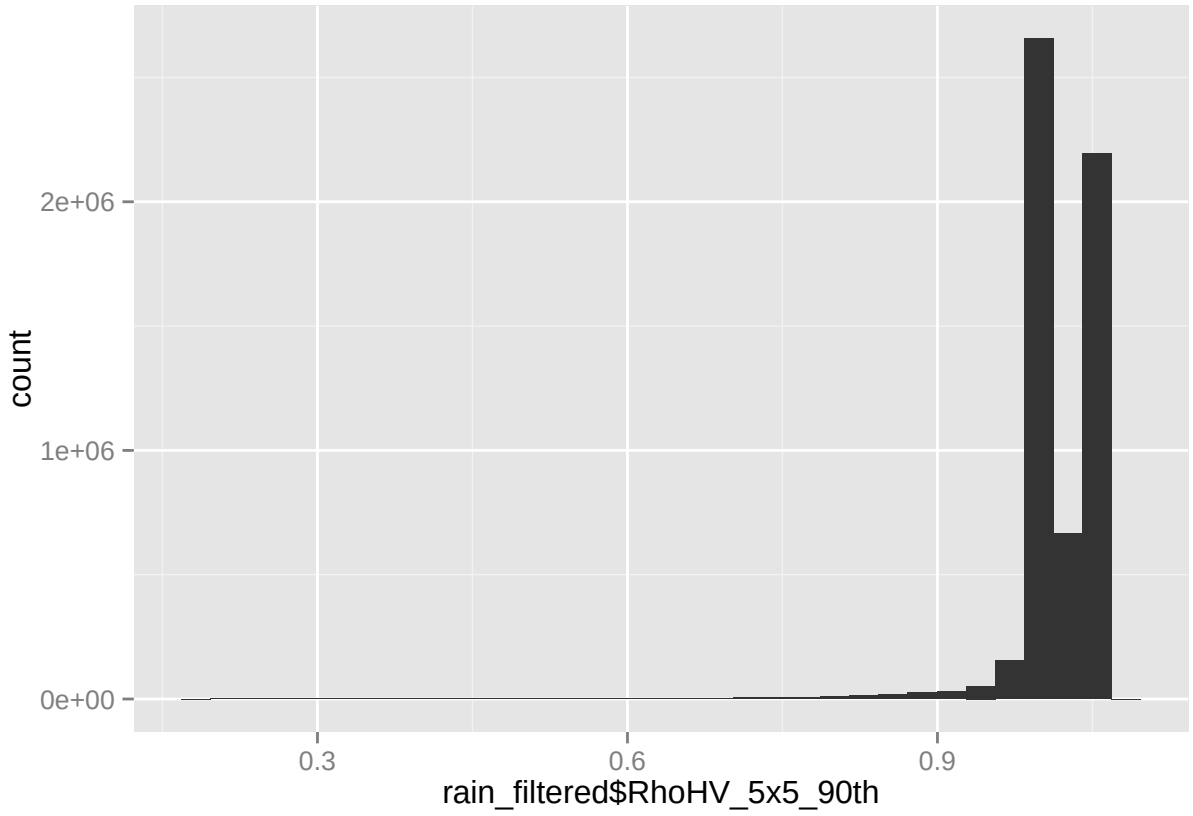

```
qplot(rain_filtered$RhoHV_5x5_10th) # very skew-left distribution
```



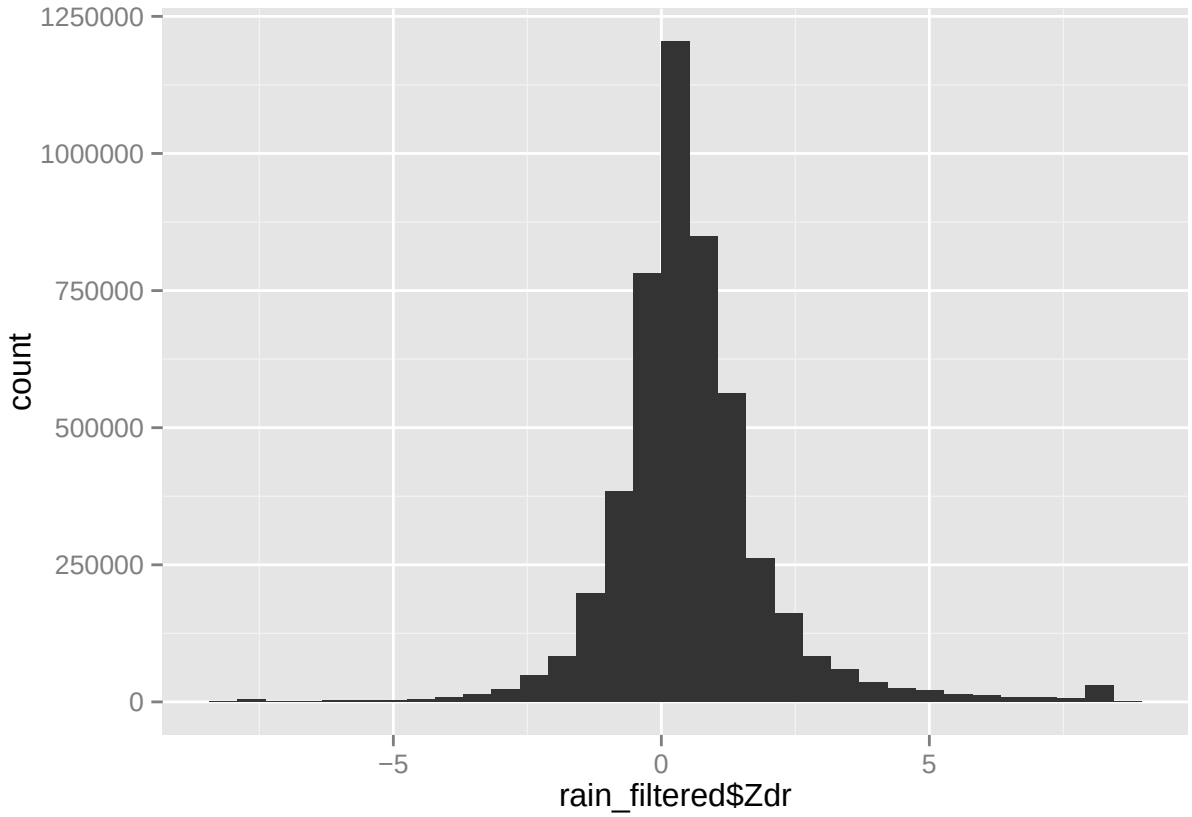

```
qplot(rain_filtered$RhoHV_5x5_50th) # very skew-left distribution
```



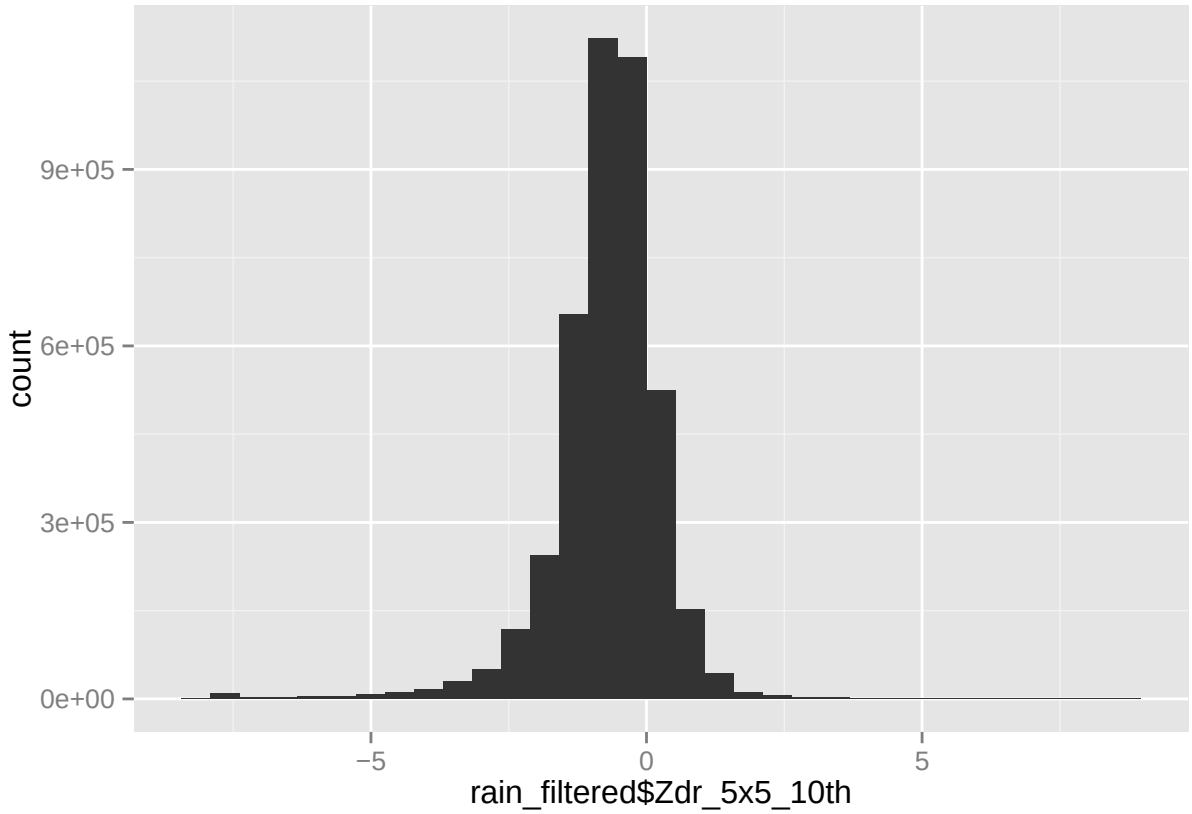
```
qplot(rain_filtered$RhoHV_5x5_90th) # very very skew-left distribution
```



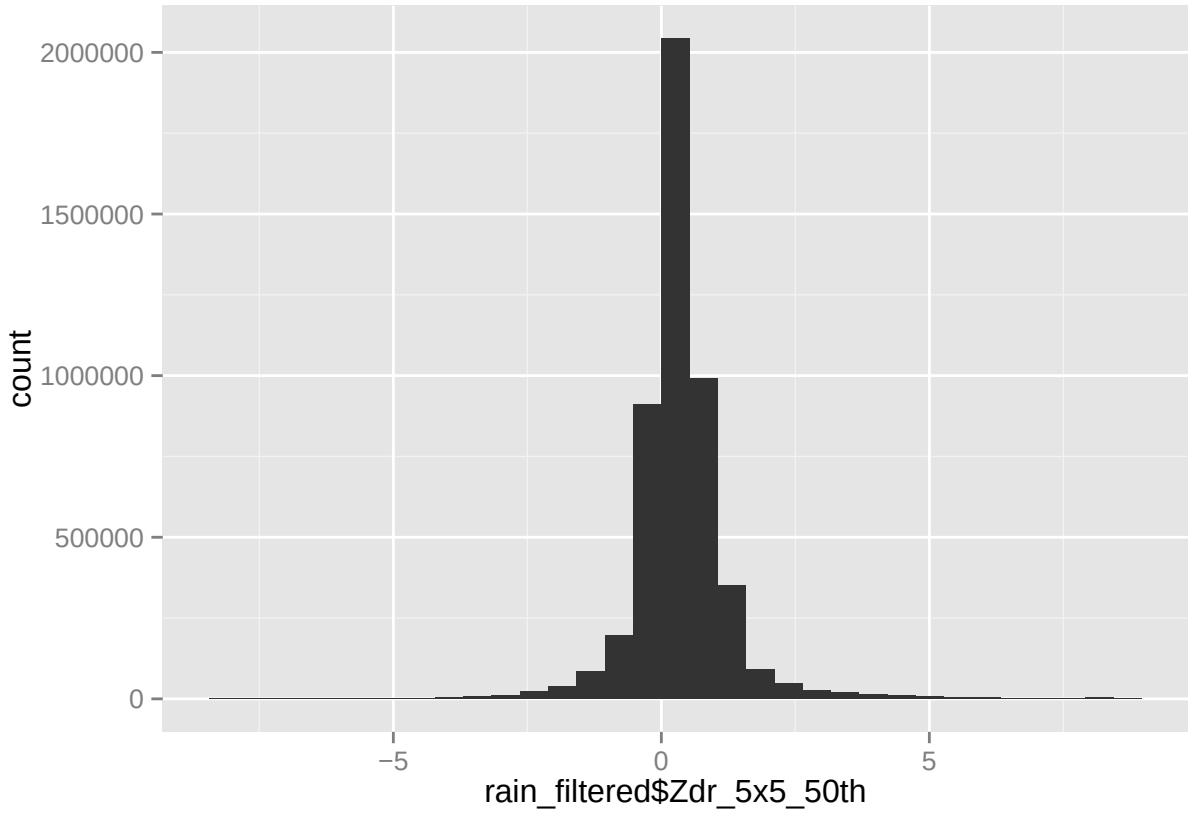

```
qplot(rain_filtered$Zdr) # normal distr
```



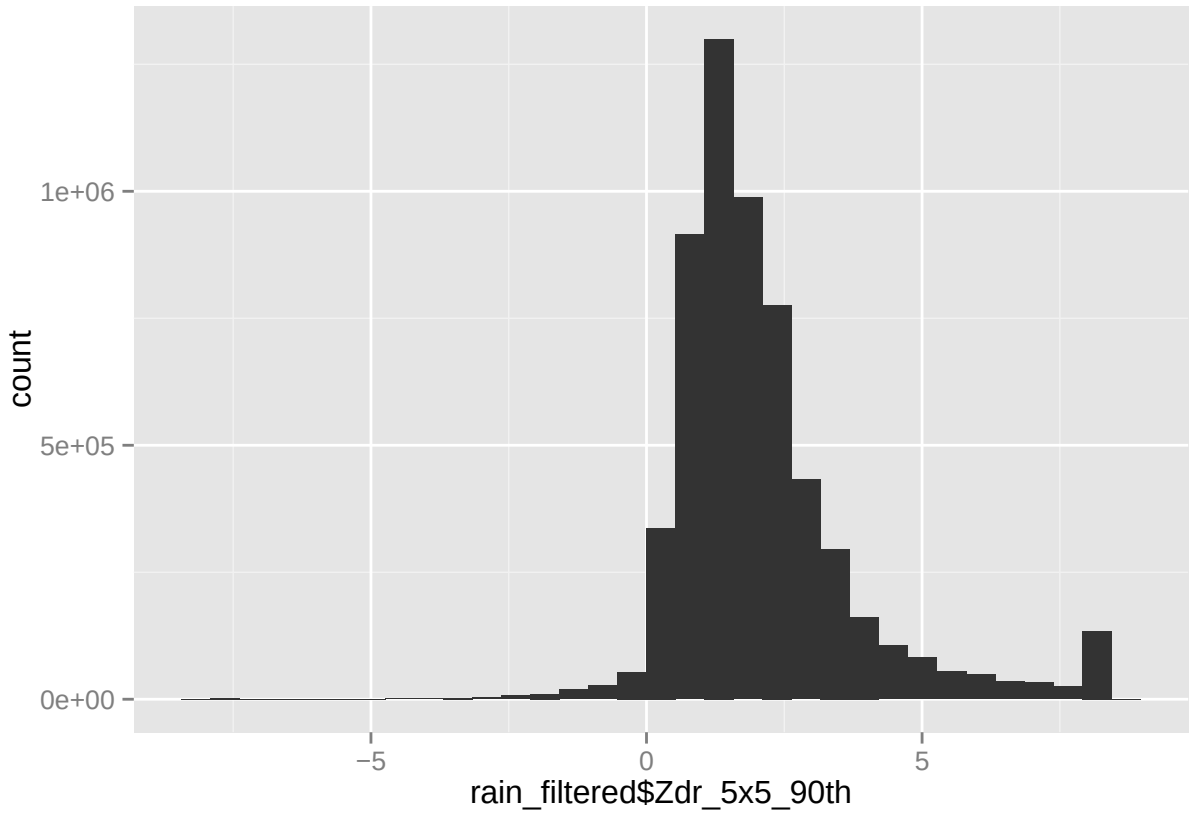

```
qplot(rain_filtered$Zdr_5x5_10th) # pretty normal dist, a bit skew-left
```



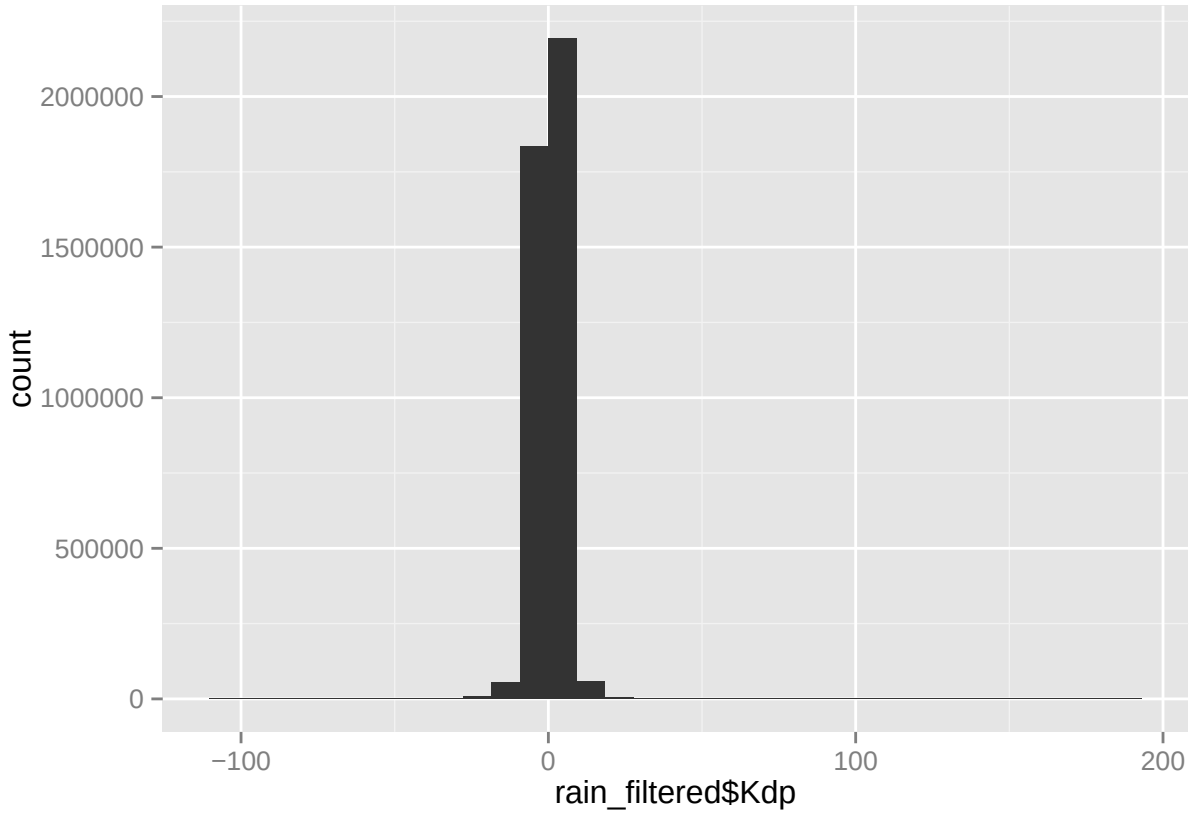
```
qplot(rain_filtered$Zdr_5x5_50th)  # narrow normal distr
```



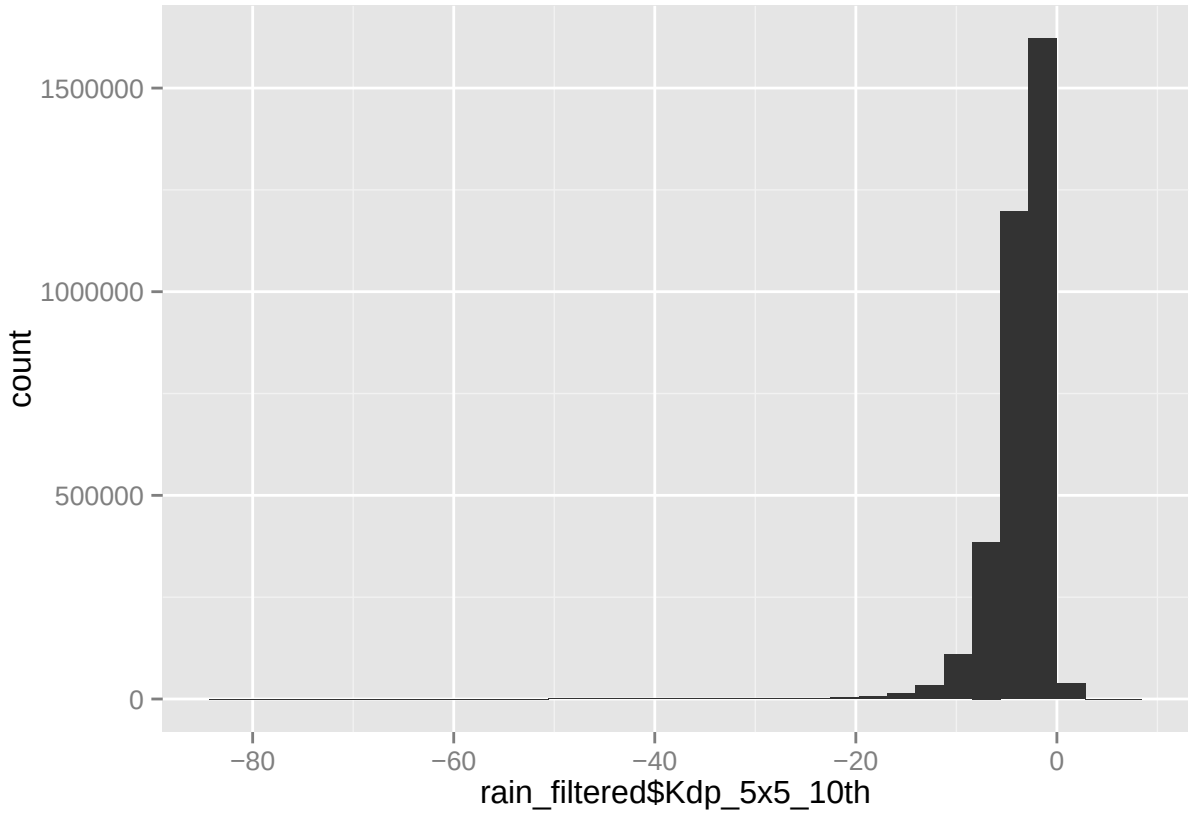

```
qplot(rain_filtered$Zdr_5x5_90th)  # interesting right-end spike -- maybe bad values??
```



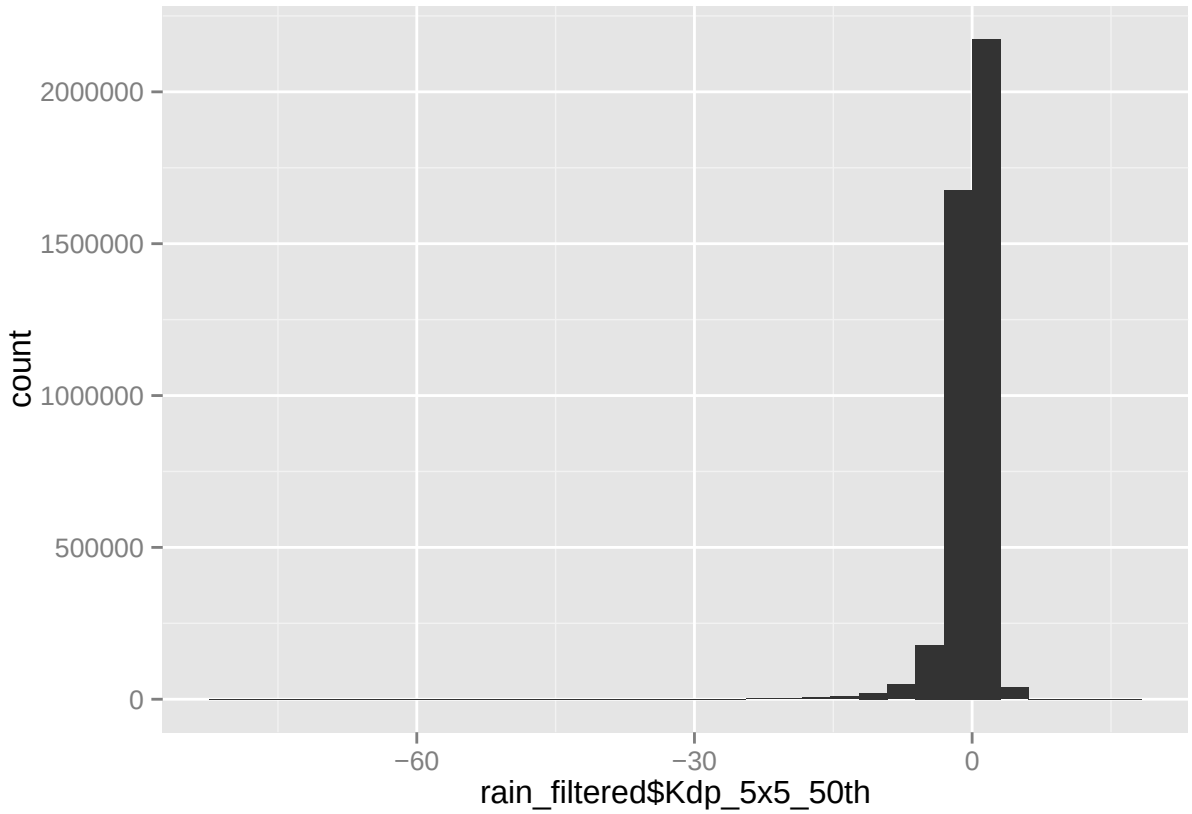

```
qplot(rain_filtered$Kdp)  # very narrow distribution with long tails
```



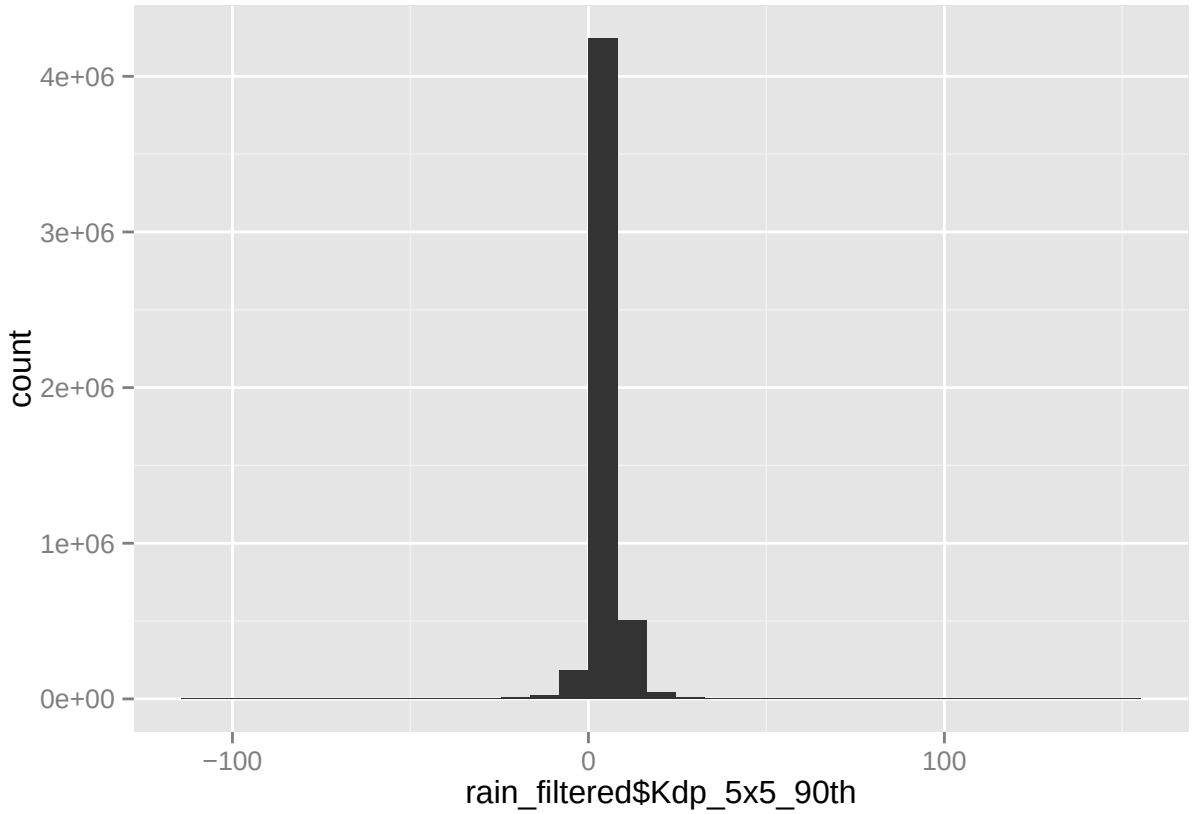
```
qplot(rain_filtered$Kdp_5x5_10th)  # left-skew distribution
```




```
qplot(rain_filtered$Kdp_5x5_50th)  # left-skew, but narrow "center"
```




```
qplot(rain_filtered$Kdp_5x5_90th)  # quite narrow dist, some very high/low values at either end though
```



So for the most part these distributions look normal and/or have some sort of skew to them. I don't think I'll need to do a lot with these data, but with the skewed variables it might be good to do log-scale transformations. I'll also need to think about interactions between variables.