Kaggle Rain Training Set - 'Expected' Distributions

Scott Mark

October 11, 2015

```
##
## Attaching package: 'dplyr'
##
## The following objects are masked from 'package:stats':
##
##     filter, lag
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
##
##
## Attaching package: 'data.table'
##
## The following objects are masked from 'package:dplyr':
##
##     between, last

## loading train from Rdata fileloading test from Rdata file
```

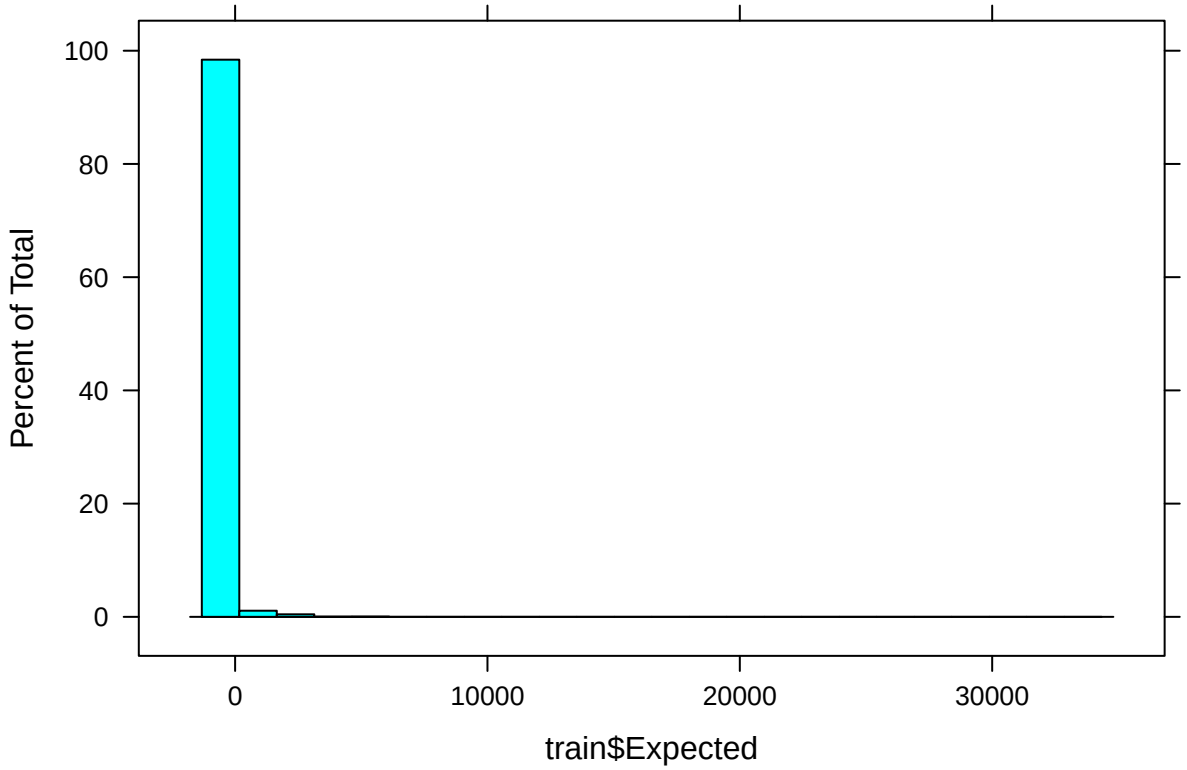
There are 2189 unique values for 'Expected' in the 8476966 rows of the training data set.

There are 0 NA rows.

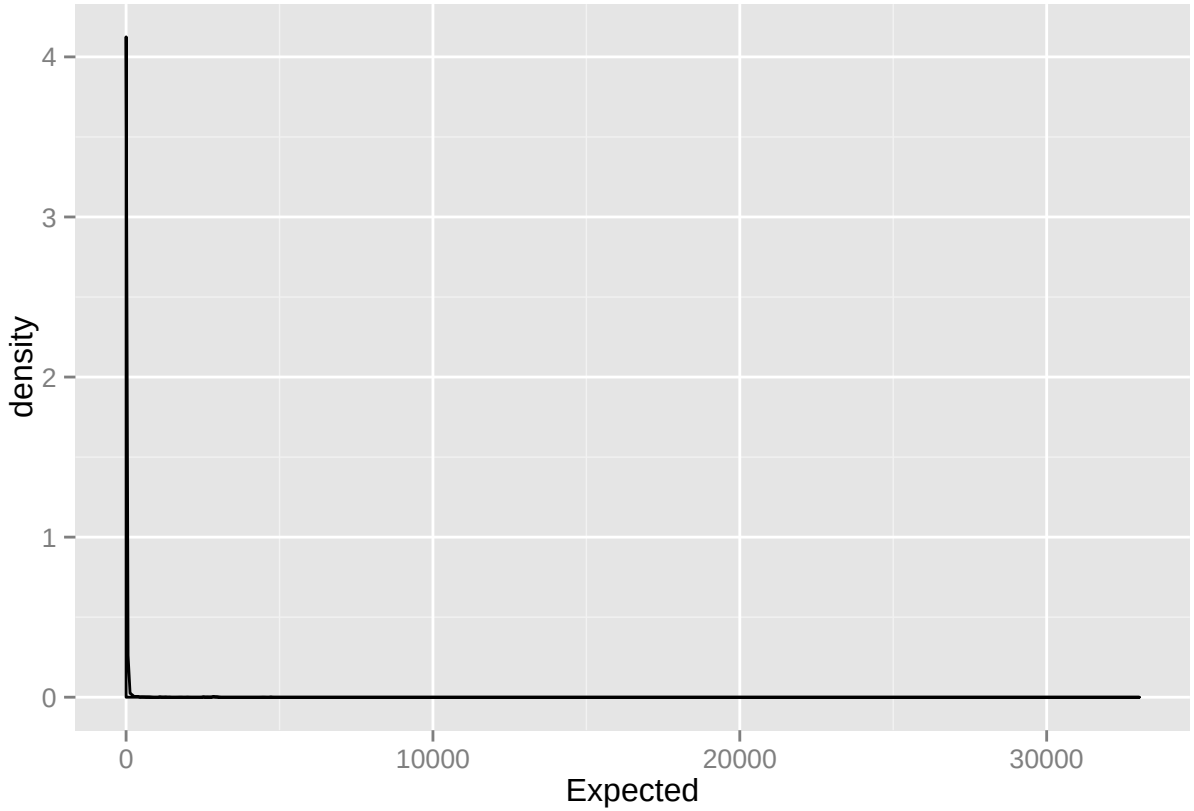
Overall Distribution

The overall distribution is heavily skewed right due to extreme outlier values.

```
plot(histogram(train$Expected))
```




```
ggplot(train, aes(x=Expected)) +
  geom_density()
```



Subset Distributions

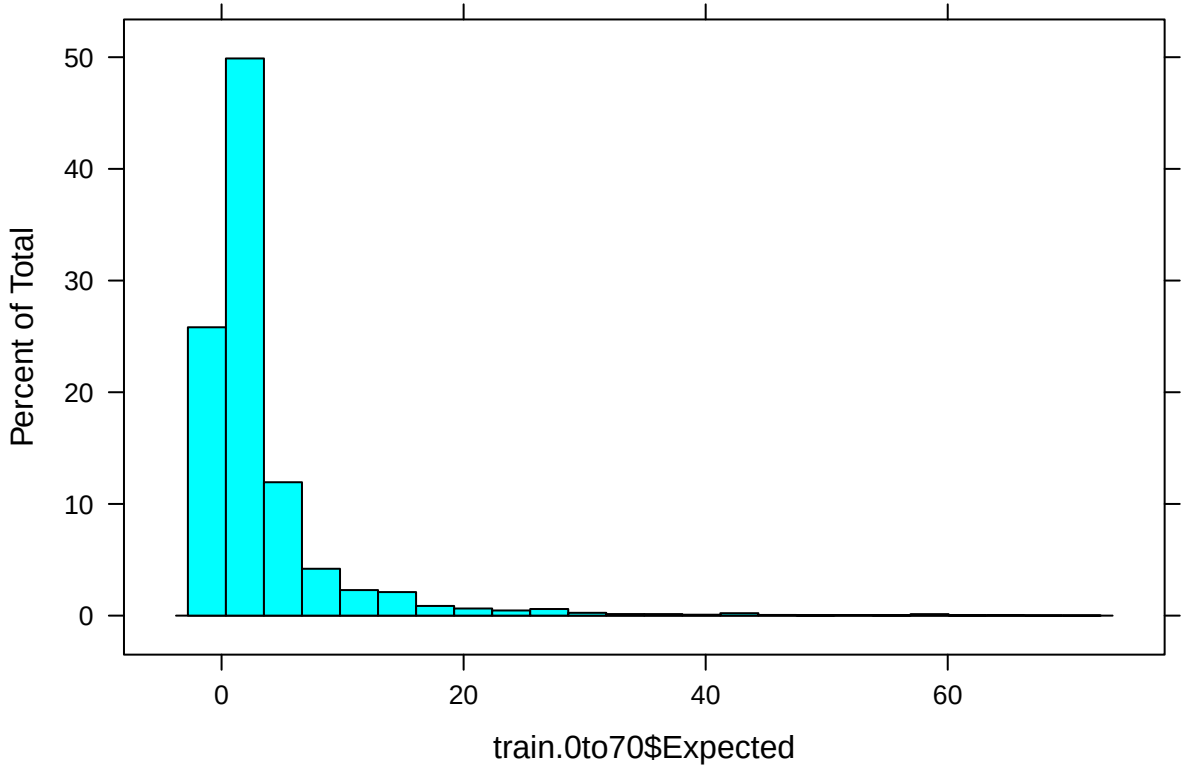
0 - 70

```
train.0to70 <- filter(train, Expected >=0, Expected <= 70)
```

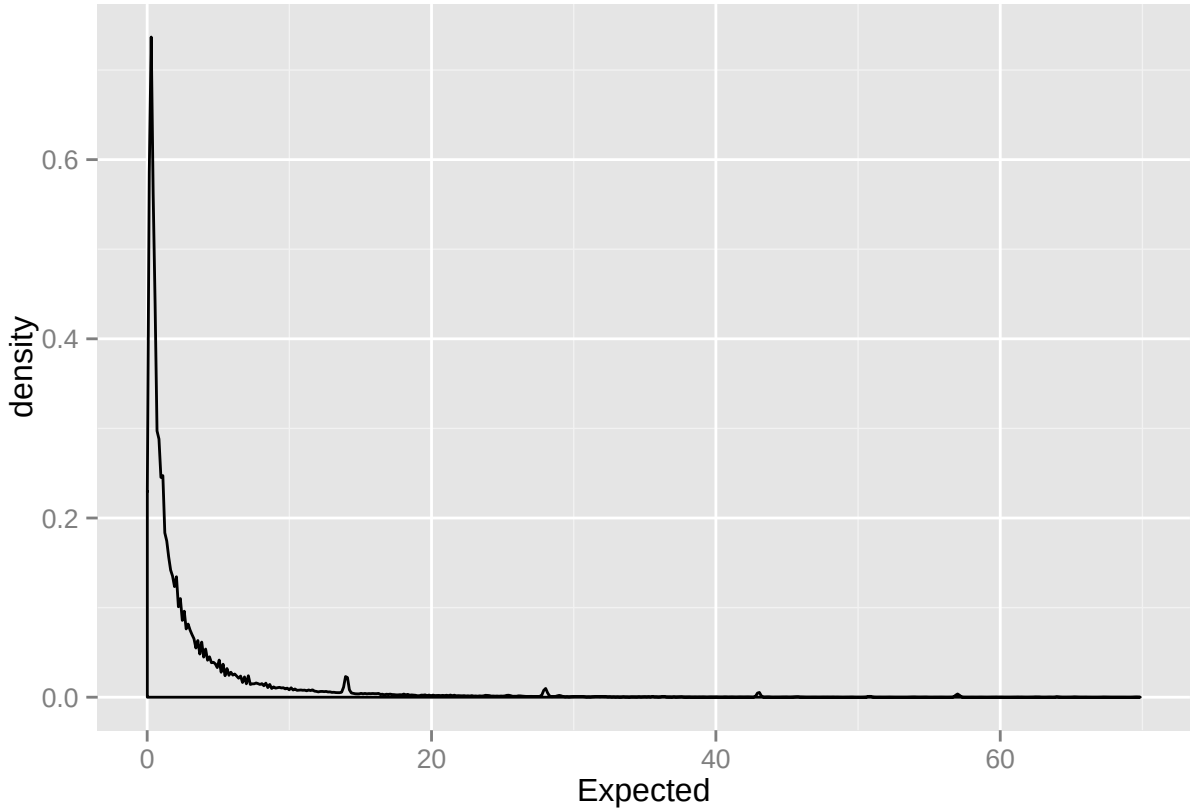
The normally expected range is 0-70mm. There are 798 unique values for 'Expected' in the 8269441 rows of this subset.

The following plots show distributions within this range.

```
plot(histogram(train.0to70$Expected))
```




```
ggplot(train.0to70, aes(x=Expected)) +
  geom_density()
```



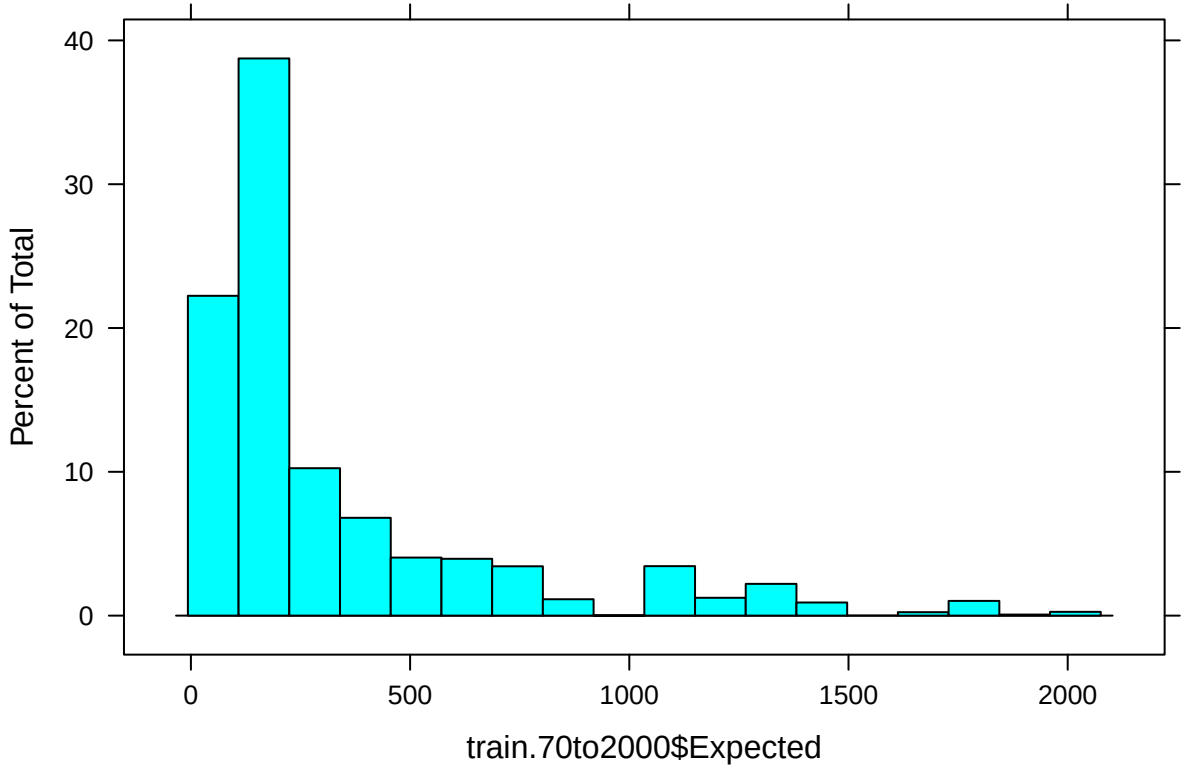
70 - 2000

```
train.70to2000 <- filter(train, Expected >=70, Expected <= 2000)
```

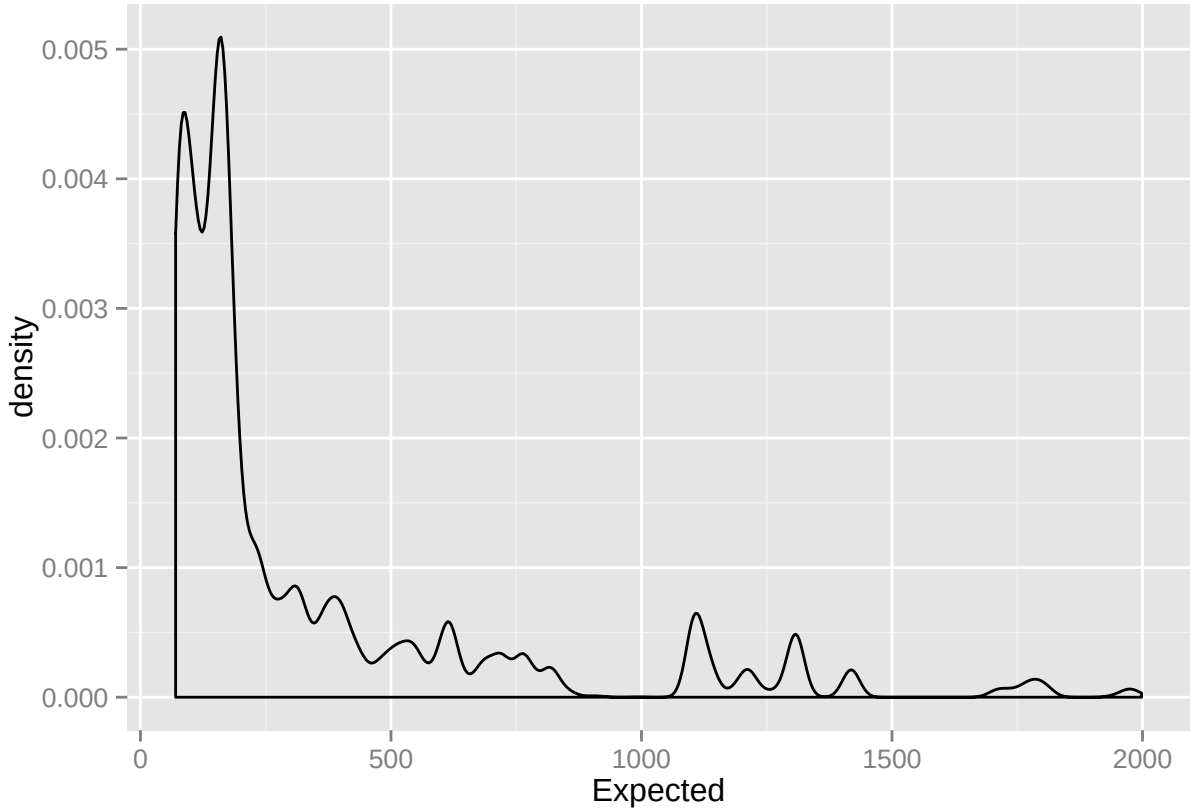
The next subset range is 70-2000mm, as there is a slight peak above 2000. There are 1008 unique values for 'Expected' in the 167555 rows of this subset.

The following plots show distributions within this range.

```
plot(histogram(train.70to2000$Expected))
```




```
ggplot(train.70to2000, aes(x=Expected)) +
  geom_density()
```



2000 - 6000

```
train.2000to6000 <- filter(train, Expected >=2000, Expected <= 6000)
```

The next subset range is 2000-6000mm. There are 370 unique values for 'Expected' in the 39836 rows of this subset.

The following table shows the distribution of values in this range.

```
table(train.2000to6000$Expected)
```

```
##
```

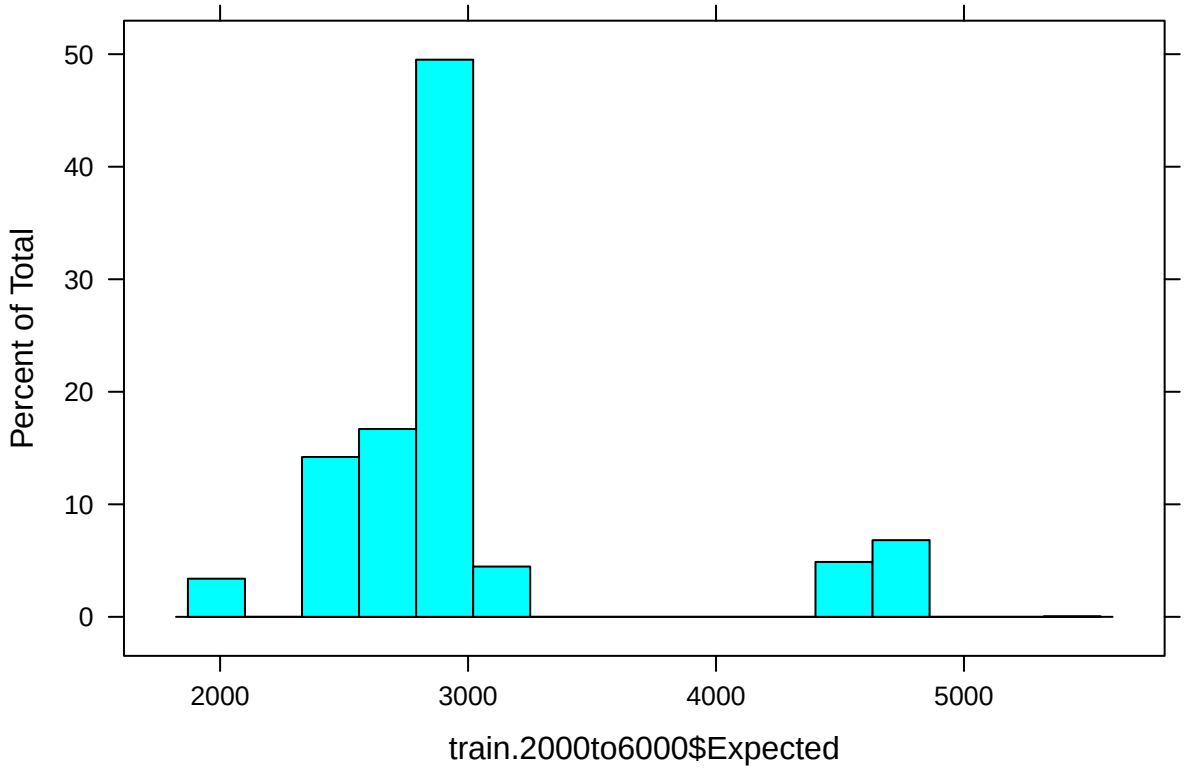
```
##  2006.347 2006.6011  2006.855 2007.3632 2008.8871 2009.9031 2013.7131
##        11        11        11        52        36       525        42
## 2013.9669 2014.2211  2016.253 2016.5071  2016.761  2020.571 2020.8251
##        12        12        11        11        11       110       202
## 2069.0852  2069.593  2079.499 2079.7532 2081.7852  2082.039  2082.801
##        64        48        10        10         6        18        54
##  2084.579 2084.8333 2381.2512 2494.7893 2495.0435 2495.2974 2495.5513
##         8        76        11        72        30       168         3
## 2503.6792 2503.9333 2504.1873 2504.9495 2505.2034 2505.4573 2511.2993
##        36        54        56        15        41       155        84
## 2511.8074 2513.0774 2513.5852 2513.8394 2514.0935 2514.3472 2514.6013
##        39        45        81        51        13        26         9
## 2515.1091 2515.6174 2517.1414 2517.6494 2519.4275 2519.6812 2519.9353
##        18        12         3       150       950       188       512
## 2523.4912 2523.7454 2531.1113 2533.1433 2534.4133 2534.9214 2537.2073
##         4       215       269        50       249         6        63
## 2537.4614 2537.7153 2538.2234 2538.4773 2538.9854 2540.0012 2540.5095
##       310       331       101       105        38        18        18
## 2541.2715 2542.7954 2543.0493 2545.8435 2546.3513 2546.6055 2546.8594
##        40        19        19       167        21        42        50
## 2547.1133 2549.9072 2550.1614 2550.4155 2550.6692 2553.4634 2553.7173
##        78       178        56        56        28        28        28
## 2554.2253 2555.4954 2555.7493 2556.0032 2558.7974 2559.0513 2559.5593
##        28        28        28        28        54        27        28
## 2559.8132 2565.6553 2565.9094 2566.6714 2573.0215 2574.0374 2574.2913
##        28        28        28       202       102        54        27
## 2575.8154 2576.3232 2623.8213 2624.0754 2625.8535 2626.1074 2627.1233
##        78       320       144       130        99        26        13
## 2627.8853 2628.3933 2628.6475 2628.9014 2629.1553 2629.4094 2629.6633
##       229        16        41        22        59        24       220
## 2629.9172 2630.4255 2630.6792 2632.2034 2632.4573 2633.7273 2633.9814
##        20        22        11        10        10        12        24
## 2634.7434 2635.5054 2638.5535 2638.8074 2639.0613 2641.3474 2641.6013
##        81        32        28        28        28        12        12
## 2642.1094 2644.3953 2644.6494 2645.4114 2645.6655 2645.9194 2649.7292
##        12        26        13        45        34         8        31
## 2649.9834 2651.7615 2652.2695 2654.0474 2657.0952 2657.3494 2658.6194
##        11        42        28        16        17        17        11
## 2658.8733 2659.3813 2659.8894 2660.6514 2661.4136 2661.6672 2665.4773
```

```
##        22        11        57       194         9       194       201
## 2667.2554 2672.5894 2672.8435 2673.0974 2678.1775 2678.4314 2683.0034
##        75        40        40       893        66        33       248
## 2686.0513 2692.1475 2692.4016 2693.9255 2694.1794 2694.4333 2694.9414
##        40       157        34        39        39        39        17
## 2695.4495 2695.7034 2695.9575 2696.2114 2696.9734 2701.7993 2702.5613
##        48       108        37        32        24        24        24
## 2703.3232 2708.4033 2708.9114 2709.4194 2712.9753 2713.4834 2714.2454
##       274        24        24        24        24        24        24
## 2718.0554 2718.3093 2718.5635 2720.5955 2721.3574 2723.8975 2724.1516
##        26        26        26       171        72        46        72
## 2725.4214 2725.6755 2725.9294 2727.4534 2727.7075 2727.9614 2729.2314
##        24        24        24        24        24        24        24
## 2729.4856 2729.7393 2731.5173 2731.7715 2732.0254 2733.8035 2734.0574
##        24        24        25        25        25        48        24
## 2735.8354 2736.0894 2736.8516 2737.1055 2796.0334 2796.5415 2798.8274
##        27        54        11        88       793       771       105
## 2805.9395 2813.3057 2823.7195 2823.9736 2824.9895 2825.4976 2826.7676
##        40        40       590       555       538        76       186
## 2827.7834 2828.7996 2831.5935 2835.6575 2837.4355 2839.7217 2840.4836
##       906       456       379        62        94        40       310
## 2845.5635 2846.0715 2846.5796 2849.8813 2850.3896 2850.8975 2851.4055
##        28        28        28       302       335        78        78
## 2852.6755 2852.9294 2853.1836 2853.4375 2858.2634 2858.7715 2861.5654
##       189      1150        34      1394       150        39       167
## 2861.8196 2862.3276 2890.7754 2892.0454 2893.3157 2894.8396 2896.1096
##       110        65        79        79        79       292        82
## 2896.6174 2897.1255 2897.3794 2897.6335 2916.6836 2919.2236 2926.5896
##       175        30        15       814       540       158        69
## 2926.8435 2931.6694 2931.9236 2932.6855 2932.9395 2937.7656 2938.0195
##        69       382       171        58       116        78        78
## 2938.7817 2945.1316 2945.3855 2959.8635 2962.9116 2969.5156 2970.0234
##       772      3139       784        68        68        68       402
## 2976.6277 2985.5176 3021.5857 3021.8396 3026.4116 3026.9197 3028.6978
##       814       128       344        64        41       126       527
## 3033.7776 3034.0317 3046.2236 3053.8445 3122.4236 4423.4126 4428.2383
##       477       102        87         4         8        11        84
## 4429.7627   4430.27 4430.5244  4431.794 4432.8105 4433.0645 4434.3345
##       101         8        16        30         1         2        15
```

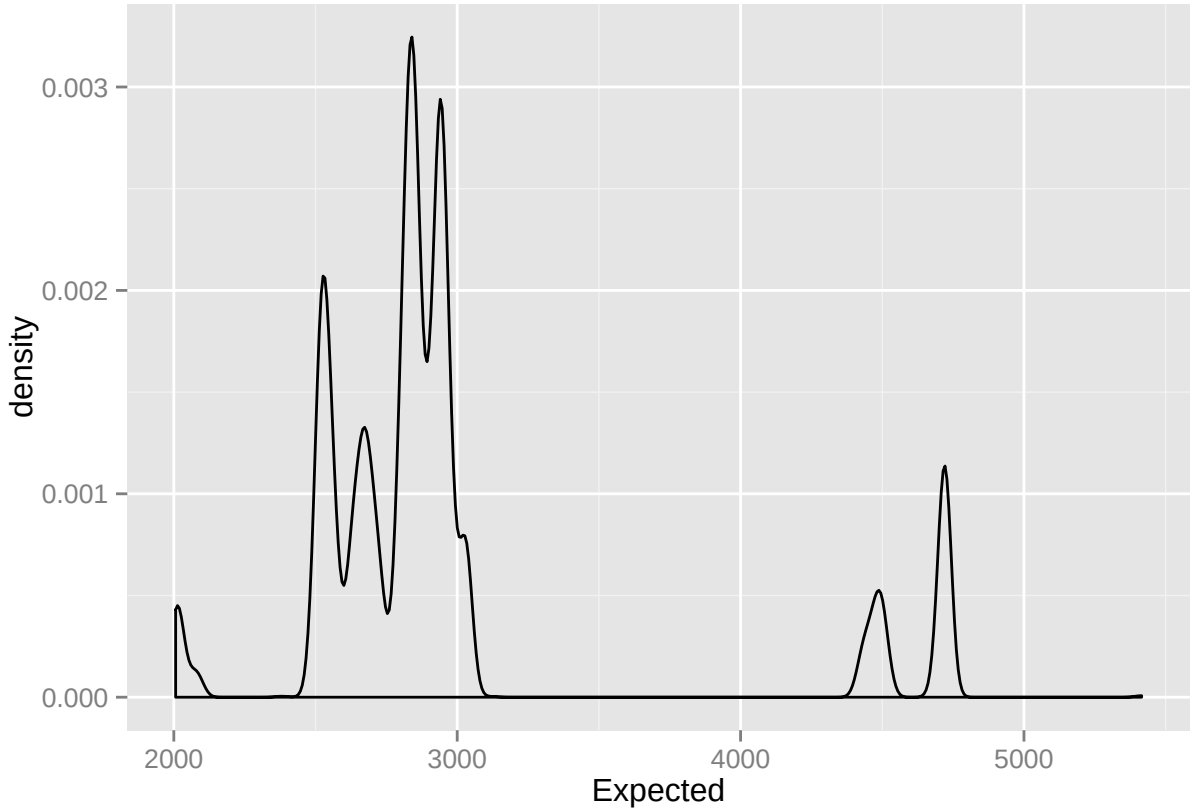
```
## 4434.8433  4435.096 4435.6045 4436.1123 4436.3677 4436.8745 4437.1284
##        26        26        20        14        14        30        11
## 4437.3823   4437.89 4438.3994 4441.4463 4441.7007 4452.1143 4452.3687
##        11        16         3         4         4        21         6
## 4453.1313 4453.3843  4453.638 4455.1626 4455.4165 4455.6704 4455.9243
##         9         9         9        11        11        11        11
##  4457.702 4457.9565 4458.2104 4459.4805 4459.7344 4461.0054 4461.2583
##        10        10        10         7        14        10        10
## 4462.2744 4462.5283 4464.3066 4464.5615 4466.3384 4467.1006 4467.3545
##         6        12        24        24        18         6         6
##  4472.434  4472.688 4472.9424 4473.1973 4473.4507 4473.7046  4473.958
##         1         2         3         9         8         3         3
## 4474.9746 4475.2285 4475.4824 4476.2446 4476.4985 4477.0063  4477.514
##         6         6        12         6         6        20        76
##  4479.038  4479.293 4480.5625 4480.8164  4481.832 4482.5947 4482.8486
##        14        28        39        13        39        14        28
## 4483.3564 4483.6104 4484.1187 4484.6274  4485.388  4485.643 4485.8965
##        16        16        16        68        39        28         4
## 4486.1504 4486.4043 4496.5645 4497.0723  4497.326 4498.8506 4499.1045
##        14         2        75        13        39        51        10
##   4500.12 4500.3745 4500.6284 4500.8833  4501.136 4501.6445 4501.8994
##        39       162        15         4         4         3        36
## 4503.1685 4503.4224 4503.6763 4504.4385 4504.6924 4505.4556 4506.2173
##         5        10         5         7        14        20         6
##   4506.47 4507.9946 4508.5024 4509.5186 4511.5503 4511.8057 4512.8203
##         2         3         4         3         8       131         8
## 4514.0913 4515.1064 4691.6377   4691.89 4720.5923 5415.2827
##         8         8        49        87      2579        17
```

The following plots show distributions within this range.

```
plot(histogram(train.2000to6000$Expected))
```




```
ggplot(filter(train, Expected >=2000, Expected <= 6000), aes(x=Expected)) +
  geom_density()
```



Over 6000

```
train.6000plus <- filter(train, Expected >= 6000)
```

The next subset range is above 6000. There are 13 unique values for 'Expected' in the 134 rows of this subset.

The following table shows the distribution of values in this range.

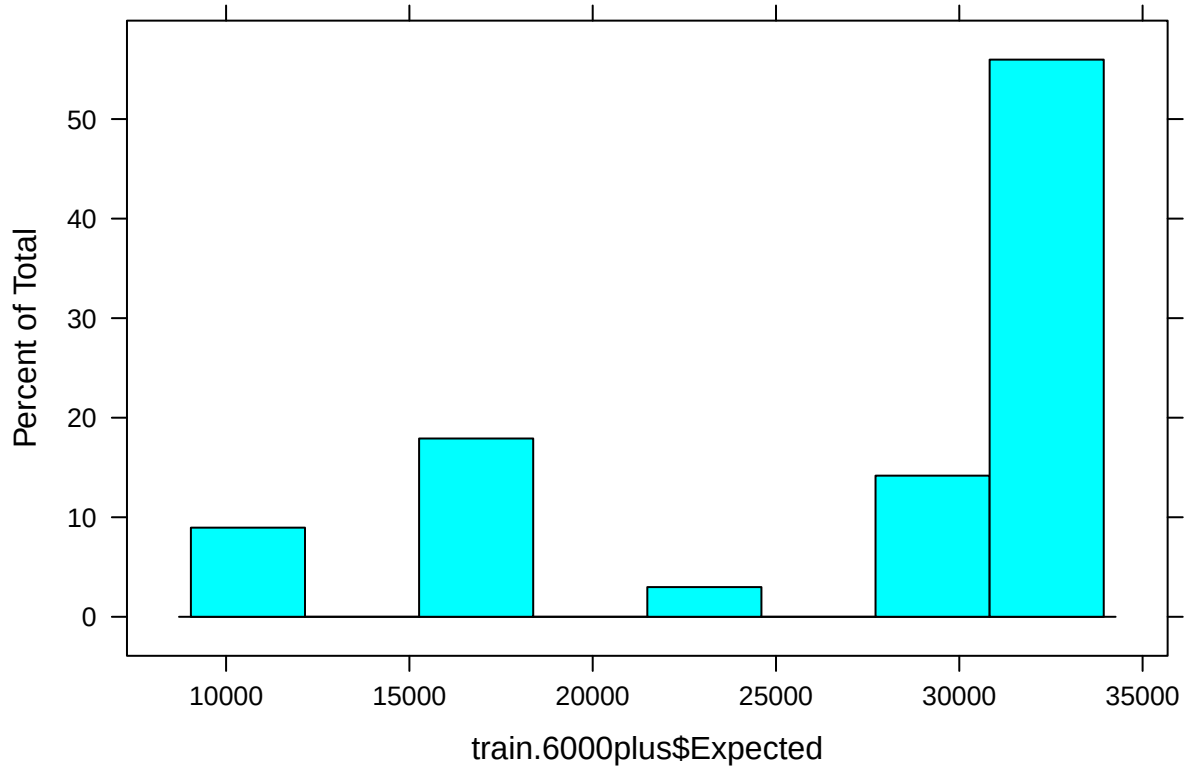
```
table(train.6000plus$Expected)
```

```
##
```

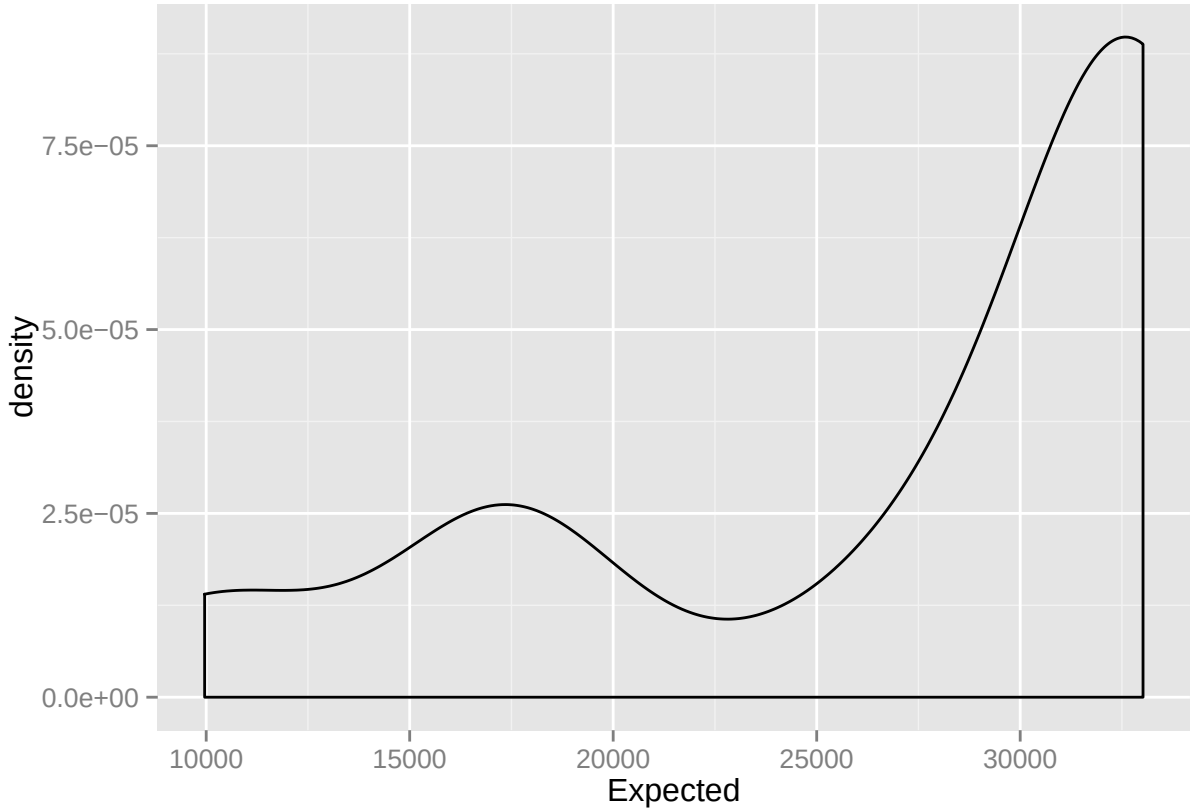
```
##  9961.887 10403.847 16553.193 18117.832 24061.943 27877.281 29697.191
##         6         6        11        13         4        13         6
## 32740.617  32779.23 32779.477 32793.445 32872.445  33017.73
##         1         1        13        15        15        30
```

The following plots show distributions within this range.

```
plot(histogram(train.6000plus$Expected))
```




```
ggplot(train.6000plus, aes(x=Expected)) +
  geom_density()
```


7.5e−05
5.0e−05
2.5e−05
0.0e+00
density
10000
15000
20000
25000
30000
Expected