Hellgate 100k — Race Analysis

brockwebb45@gmail.com

February 12, 2015

Synopsis

The Hellgate 100k is a tough, but rewarding race directed by Dr. David Horton. Taking place around the second week of December in the mountains just north of Roanoke Virginia, this race is known for challenging and unpredictable weather that make each running a unique experience. The goal of this analysis was to look at the data and see if there were any interesting patterns or results to investigate. Also, the basic stats could be used to create a nice info-graphic highlighting Hellgate.

Methodology

Please note: all the data files used and complete source code is available from the github repository at: https://github.com/brockwebb/hellgate_100k_statistical_analysis

All race time data and weather information was obtained through the referenced web sources, cleaned, and transformed into comma separated value (csv) format for ease in any analysis application. The weather data used was from the Roanoke Regional Airport (Weather Data). No data existed for temperatures at higher elevations like Headforemost Mountain at mile 24/Aid Station 4, which is known as the coldest section of the course when runners reach it before dawn. The Airport data was believed to be the best and most reliable source for this information due to its criticality in aviation safety. Taking windchill in account was done using the average observed wind speed and the lowest observed temperature as an estimate of what was felt on the course, albeit Hellgate has higher elevations, receives more wind exposure at the top of the mountains, and was probably colder.

Normality tests on the data distributions were conducted by applying three methods: Shapiro-Wilk test, Anderson-Darling test, visual inspection of the Q-Q plot. The non-parametric Mann-Whitney U Test was used for all non-normal data comparisons. In all tests, significance was the standard accepted p-value of $p < 0.05$.

Linear models were used to show best fit whenever appropriate. A second order polynomial was used in one case as the points demonstrated a curved pattern. For the predictive modeling of Hellgate times based on the Beast Series, the step() function was used to attempt the best prediction from the data. Regardless of the R^2 value obtained from the model fit, regression analysis was used to investigate goodness of fit. Regression analysis included an examination of the resulting residual and Q-Q plots for any patterns that would suggest complications or problems with any data fitting models used.

Results and Conclusions

All charts and in-depth analysis are found in the Appendix. This section provides a summation of the most salient results and conclusions.

Male vs. Female Runners

The age distribution of male and female runners was equivalent (p=0.4879). Overall, males finished faster than females (p=0.00015) with an average time of 15:33:47 to 16:05:47 for the women. The most surprising

result was that there were no significant differences between male and female finishing times for the first five years(p=0.2482). After that time, males became significantly faster whereas female times seem to stay similar to previous years (p<0.00016).

2008+ Improvement in Male Finish Times

The reason for this improvement or trend is still unknown. The improvement effect was also noted in this analysis: "Increase in finishers and improvement of performance of masters runners in the Marathon des Sables" (Jampen SC, Knechtle B, Rust CA, Lepers R & Rosemann T, 2013). Books like *Born to Run* by Christopher McDougall was first published in May 2009, drawing a broader audience into the sport, as well as barefoot running. According to finishing stats from Ultrarunning Magazine, Ultra finishes went from 30,789 in 2008 to 69,573 in 2013 (UltraRunning Magazine, 2013). Overall, the effect may be attributed to an increase in popularity of the sport leading to increased participation that uncovered a wealth of undiscovered talent.

Several factors were investigated, including looking at the performance of veterans of the race, use of the race committee in applicant screening, and whether or not the Beast Series, which began in 2008, was a factor. The Beast Series did not explain the improvement, but had power in predicting the Hellgate finishing time, covered later in this section.

The question of whether or not the new race registration procedure of screening applicants is making a difference is undecided because the trend began three years before the introduction of it. The race committee may be stabilizing the finish percentage or improving the overall time, but the effect requires further study when more data are available.

Finish Rate

The early years had a much lower finish rate than the later years. The trend observed implied a "learning curve" and fit a second order polynomial trend-line nicely. Explanations of this learning curve might be due to knowledge of the race, race reports, and potentially the race committee screening procedure. As the improvement in finish times suggests, better prepared runners are showing up and future races are predicted to finish in the low 80's.

Temperature and Finish Rate

The warmest start was 2004, but factoring in windchill, 2013 was the warmest year to date. The coldest years were 2007, tied with 2009 (20 degrees F). Factoring in windchill, 2007 was the coldest around 9 degrees F. Temperature appeared to correlate with the finishing percentage, but closer examination of the residual plots revealed non-linear behavior, meaning some other factor was not accounted for in the result.

Course Conditions

Runners experienced clear trail conditions six times, ice two time, and snow-ice four times. Course conditions had no significant impact on finishing time (all p>0.05). Overall, the wind direction was to the East five times, North-West three times.

Beast Series

Beast series runners were not significantly different than their peers as far as finish time are concerned. Only one year, 2011, really looked visually different from the box plots. By Removing the few outliers in 2011, Beast runners finished slower than their peers (p=.0325).

Past performance in the Beast Series was examined to determine how well it could predict a runners finish time at Hellgate. The following equation was found to provide a reasonable estimate of a runner's finish time:

*Hellgate time (predicted) = (-2.948e+04) + (1.897)MMTR.time + (5.684e-01)GS100.time + (-1.128e-05)*MMTR.time$GS100.time$

Where: MMTR.time and GS100.time are the finish time from each race, in decimal based hours (e.g. h:m:s 16:30:00 = 16.5 hours).

This equation has an adjusted Rˆ2 = .08075 and has decent looking residual plots, confirming the model's validity as a good predictor.

Future Study and Investigation

Future study might include a broader, more in depth look at a runner's experience and preparation before Hellgate. More data and insight into how the Race Committee selects runners from the applicant pool would be helpful in determining its effectiveness. It would be interesting to see if the "learning curve" effect exists for other races and how long it takes to stabilize around a given level. When more data are available, the Hellgate prediction model should be evaluated to measure its robustness. Lastly, it would be prudent to check other race performance data to see if the 2008/2009 time frame kicked off an increase in overall performance throughout the sport.

Appendix

The sections below execute are constructed though execution of all the R code and were the basis of the above report.

Environment Setup

Global Options

Global options set options for general formatting, including suppressing messages and warnings from the code. There are many messages produced that add length to the document and distract from readability. Also, turning on/off all code, chart sizes, etc. to display results is possible here too.

Data libraries

The following libraries are used in this analysis: 1. ggplot2 – graphics/charts 2. grid – layout of charts for display 3. lubridate – handling date/time 4. knitr – knitr global options for output/file build 5. nortest – normality testing
6. plyr – summarization/aggregation

Global Functions

One global function is used to return the equation and rˆ2 value that is generated by the linear model for display on graphs using ggplot (Regression Equation)

Loading the Data Files

All the data files used, including this R-Markdown file with complete source code is available from the github repository at: https://github.com/brockwebb/hellgate_100k_statistical_analysis

Information on which runners actually started, did not finish (DNF), and where they dropped was not available. This may have aided even more perspective into the complex challenges this race presents and where the major obstacles are.

Basic Stats:

Below are the basic stats on the race:

Time/Age Records

Here are the male and female finishers' fastest/slowest times and youngest/oldest ages:

Total finishes:

```
## [1] 1032
```

```
##                 stat      male    female
## 1   Total Finishes       853       179
## 2   Fastest Finish 10:45:49 12:23:40
## 3   Longest Finish 18:55:00 18:55:54
## 4     Average Time 15:33:47 16:05:47
## 5 Youngest Finish        19        21
## 6   Oldest Finish        66        62
## 7     Average Age      39.9      39.4
```

Runners by state

Looking at the distribution of runners from each state, top five states, totaling 71.1 percent of all finishes are from Virginia, Pennsylvania, Maryland, North Carolina, and Ohio:

```
##     state runner.count percent.by.state
## 34    VA          468             45.3
## 29    PA           86              8.3
## 16    MD           76              7.4
## 20    NC           65              6.3
## 25    OH           39              3.8
```

```
## [1] 71.1
```

Finish Percentage by Year

The number of starters has increased from 71 (2003) to 148 (2014). The Number of finishers has also gone from 44 (2003) to 122 (2014). The best finish percentage was 88.1% in 2010. Also show are the "First Timers" – those whose first Hellgate finished in a success. Interestingly enough, there was a noticeable jump in 2010 in

the total number of new runners. The fact that the highest finishing percentage and the largest number of new runners occurred in the same year is probably a coincidence, as there is not enough data to say otherwise.
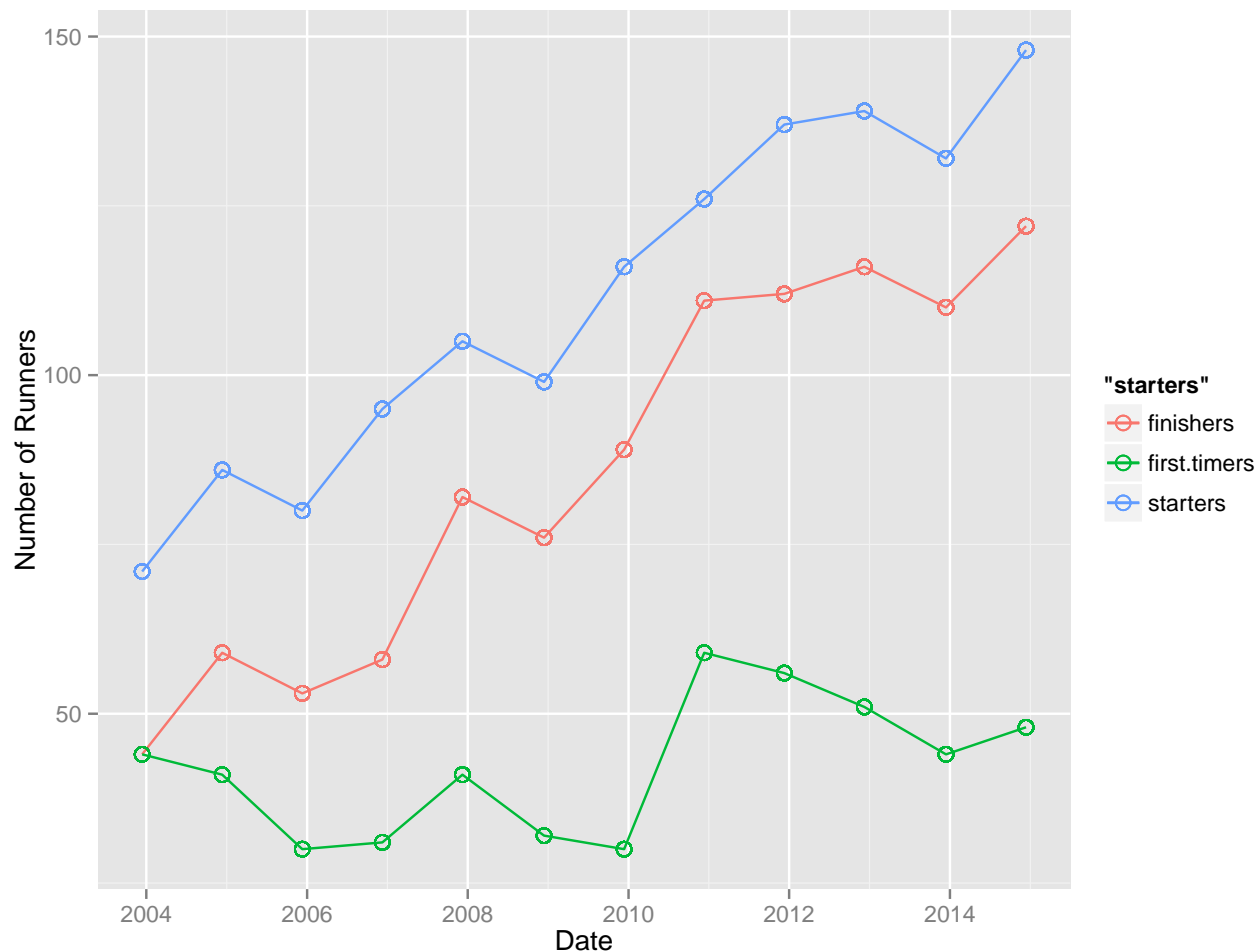
In 2011, the race registration format changed. Entry criteria required that each runner list the races they have done to "prove" they are able to complete Hellgate. A race committee then reviewed the applications to ensure runners had a good chance at finishing. It would appear that a "learning curve" with finishing may have occurred anyways, as more runners gained the benefit of race reports and prior Hellgate experience. Future race predictions would indicate a finish percentage in the low 80's, and the effect of the Race Committee might have stabilized the success rate of Hellgate in this area. This "learning curve" is represented by the smoothed line in the plot below:

Weather Effects

Hellgate can have some interesting weather, and the course conditions are dependent on many factors leading up to the race (snowfall, rain, cold/heat waves, etc). Weather data was obtained from the U.S. National Oceanic and Atmosphere Administration's (NOAA) National Climatic Data Center as measured from the Roanoke Regional Airport (Weather Data). No data existed for temperatures at higher elevations like Headforemost Mountain at mile 24/Aid Station 4, which is known as the coldest section of the course when runners reach it before dawn.

The Airport data was chosen because wind speed/direction is incredibly important for airplanes and was believed to be the best/most reliable source for this information. This was important in calculating wind chill by using the formula from the National Weather Service (Windchill Calculation). It should be noted that the formula was only applied when average wind speeds exceeded the 3 miles per hour (mph) threshold where it is considered valid (Windchill Calculation).

The course conditions were recovered from the race reports, but mainly from the Race Director's (Dr. David Horton) account of the conditions (Extreme Ultra Running).

Temperature

Plotting the both the Temperature and the Wind Chill using the average wind speed and coldest observed temperature at Roanoke Regional Airport. The warmest start was 2004, but factoring in windchill, 2013 was

the warmest year to date. The coldest years were 2007, tied with 2009 (20 degrees F). Factoring in windchill, 2007 was the coldest around 9 degrees F.
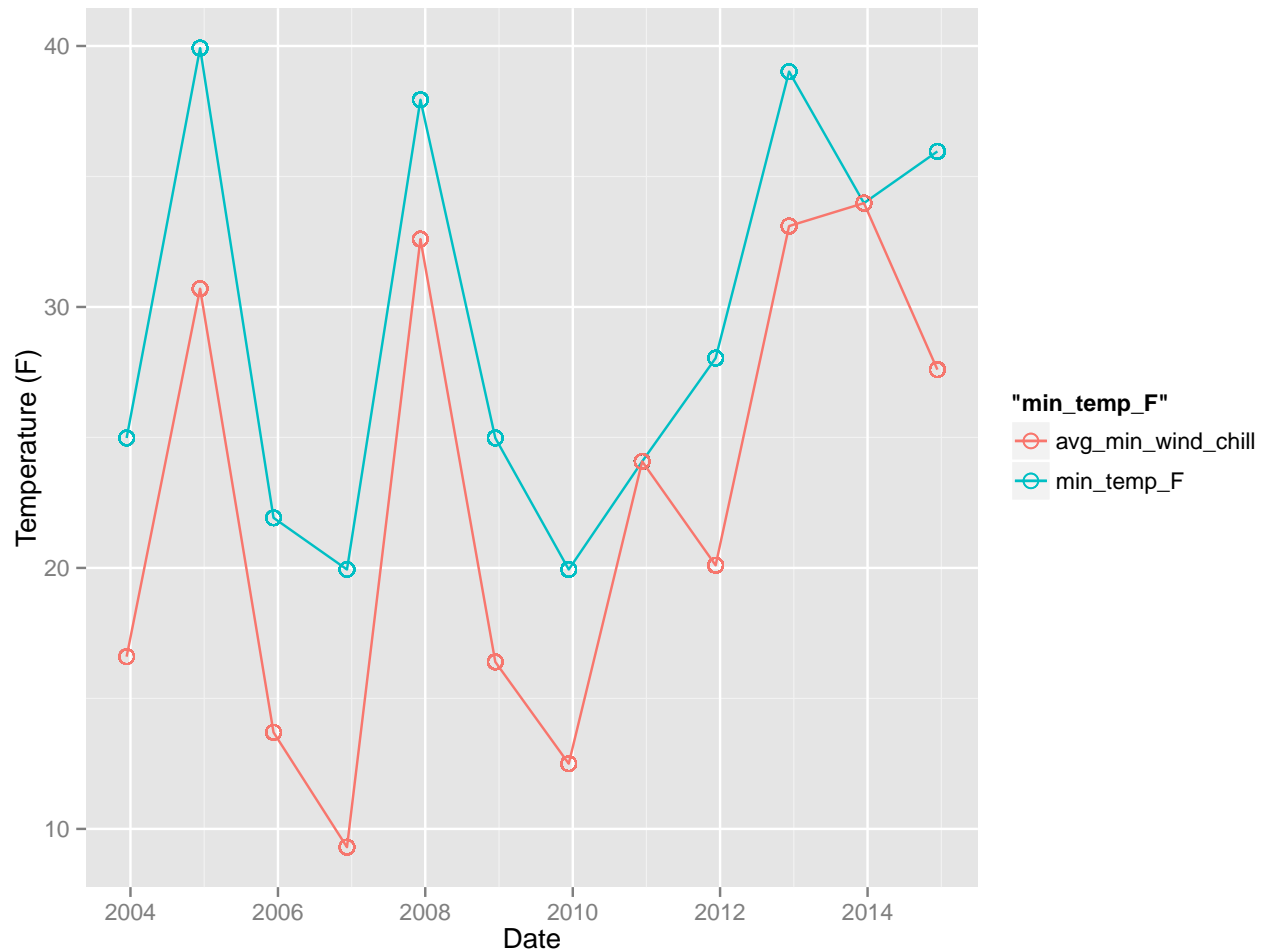

Wind Direction

Historical "wind direction" was not found in the data set so wind direction was determined by using the five minute wind gust direction as an estimate for the overall direction the air system was moving. The five minute wind gust is the highest sustained wind for a five minute period during the observation time. The data set contains direction in the form of degrees in a circular reference where east is zero degrees, north is ninety degrees, and so forth. Wind blowing in the south direction was the most common with five occurrences, north west the second most common with three:
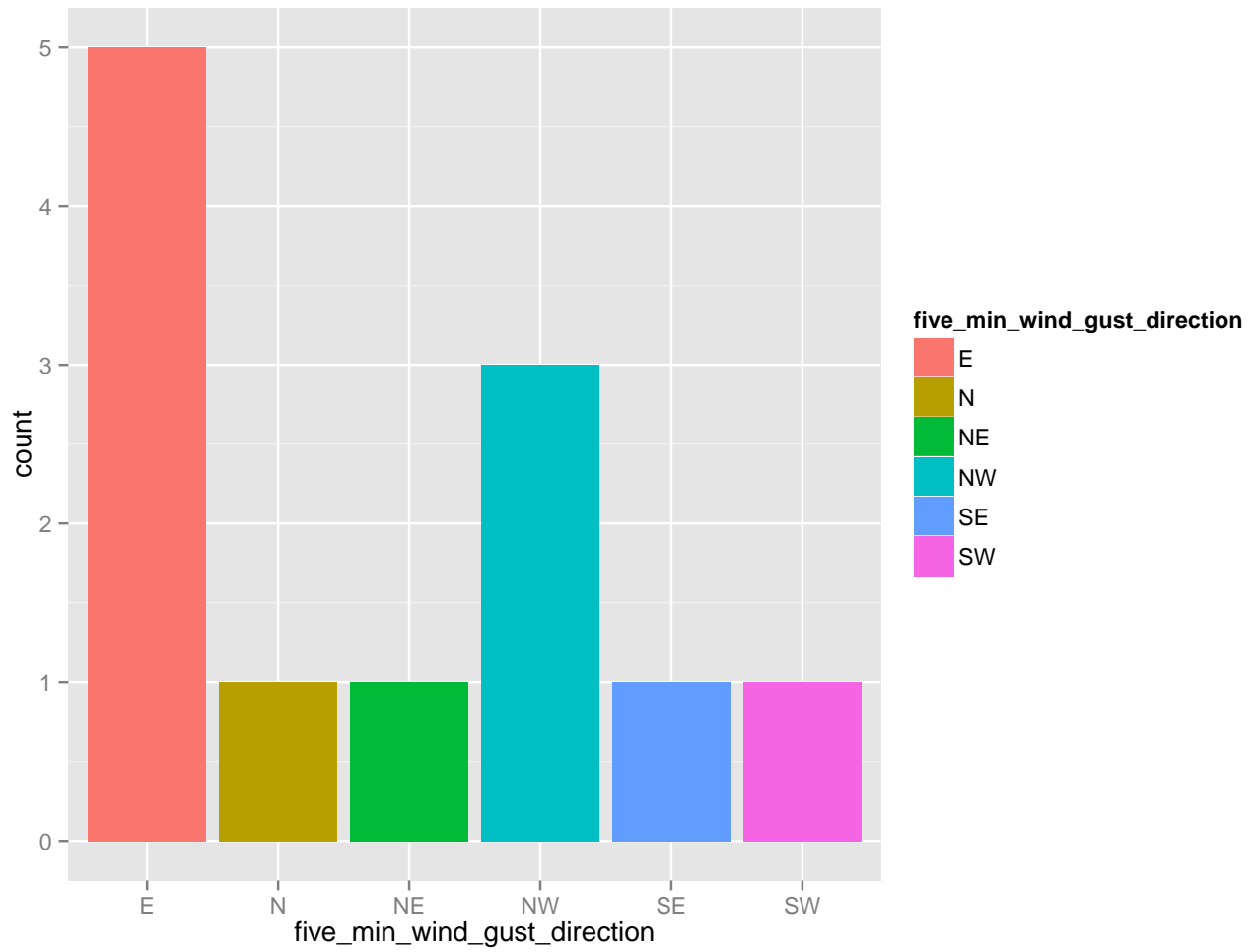
Course Conditions:

Most of the time, the trails were clear. However, some years had ice or a mixture of snow and ice on the ground:
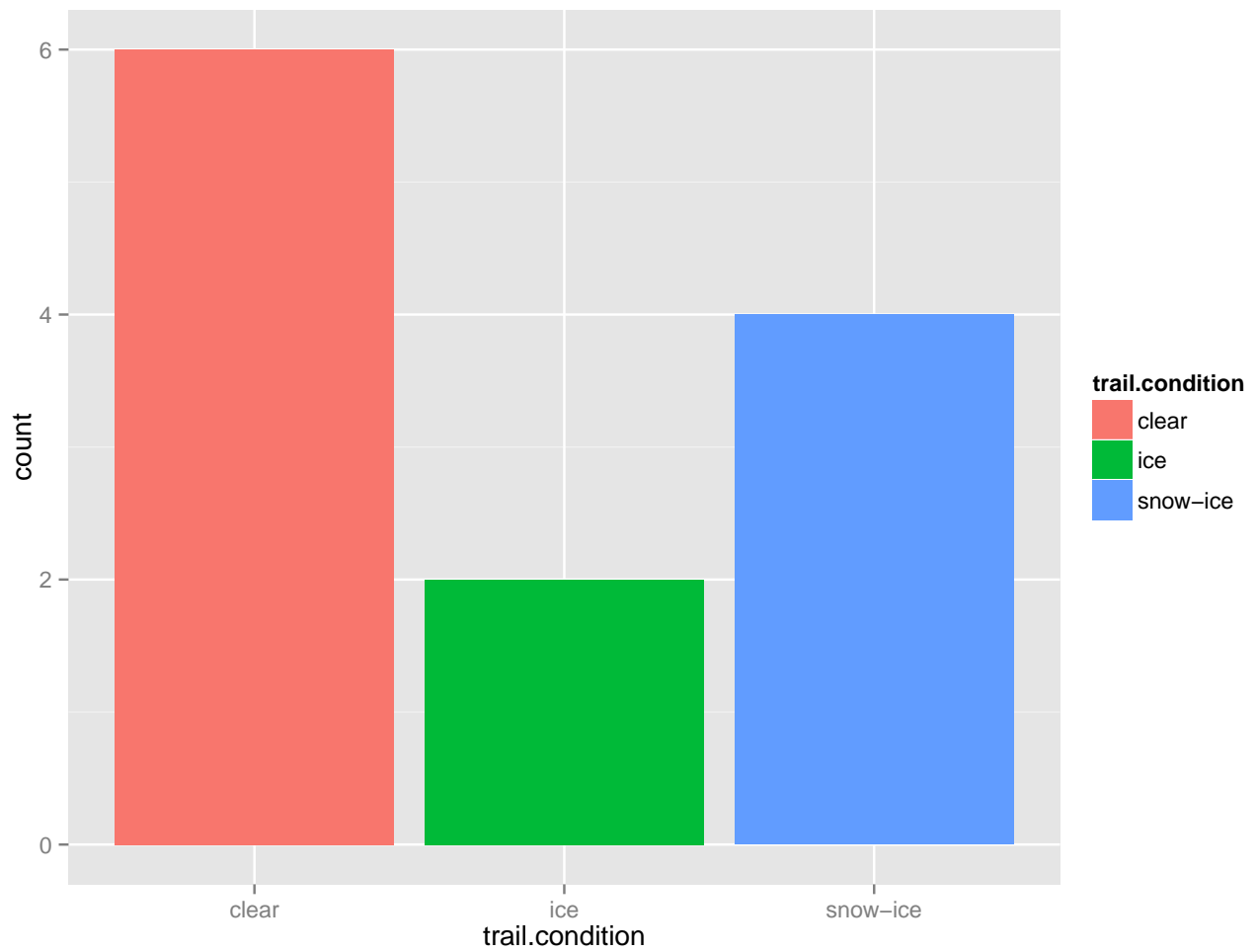
Statistical analysis

Plot of male versus female times

Looking at the overall finishing times distributions by gender:
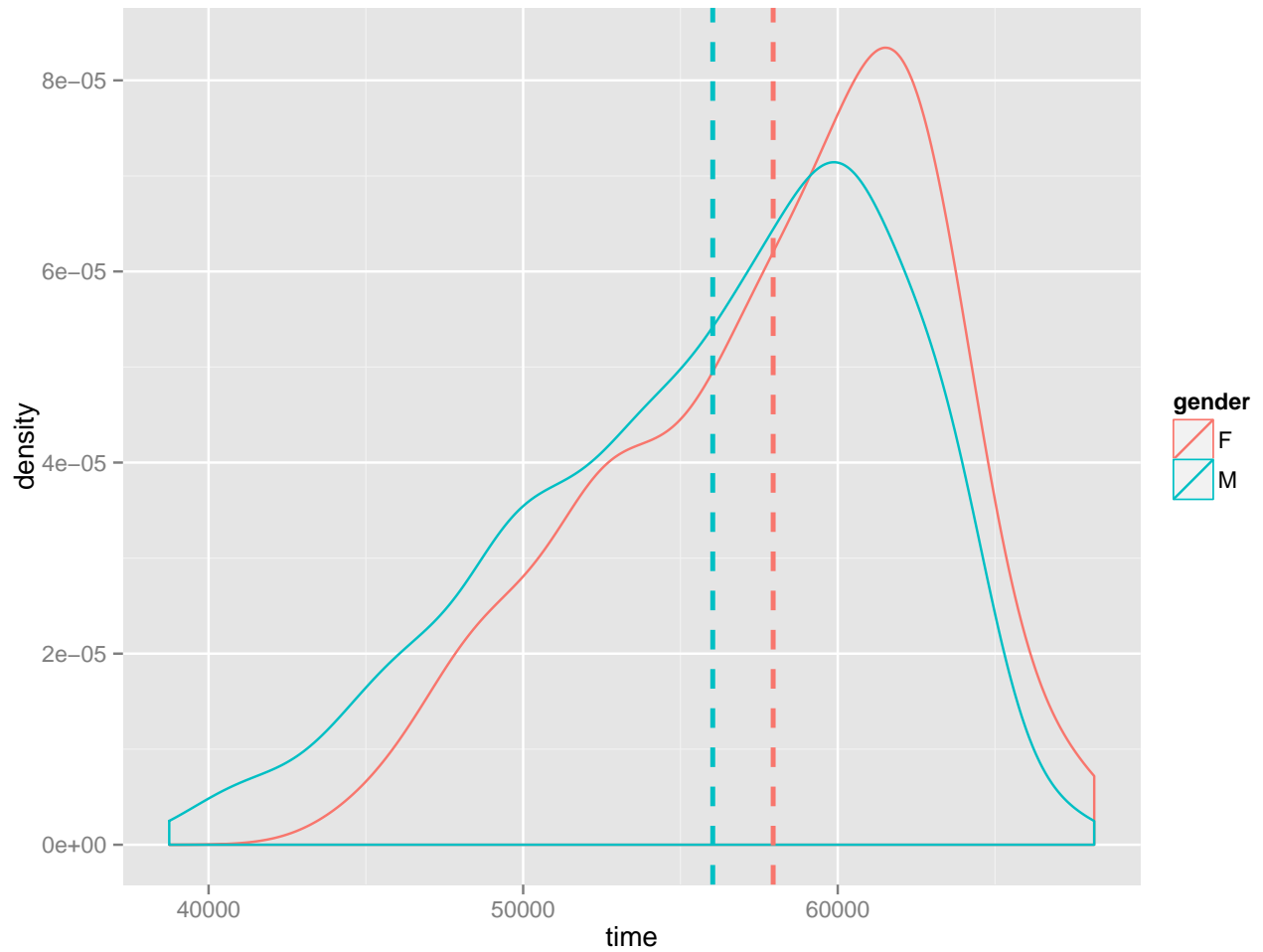
The time distributions are similar, but the median time is different. The shape is not too normal looking. Much of the data has this type of skew, mainly because a larger portion of the runners finish much later in the race. To prove non-normality, applying the Shapiro-Wilk and Anderson-Darling tests and confirming Q-Q plot produces the following for the Men's times:
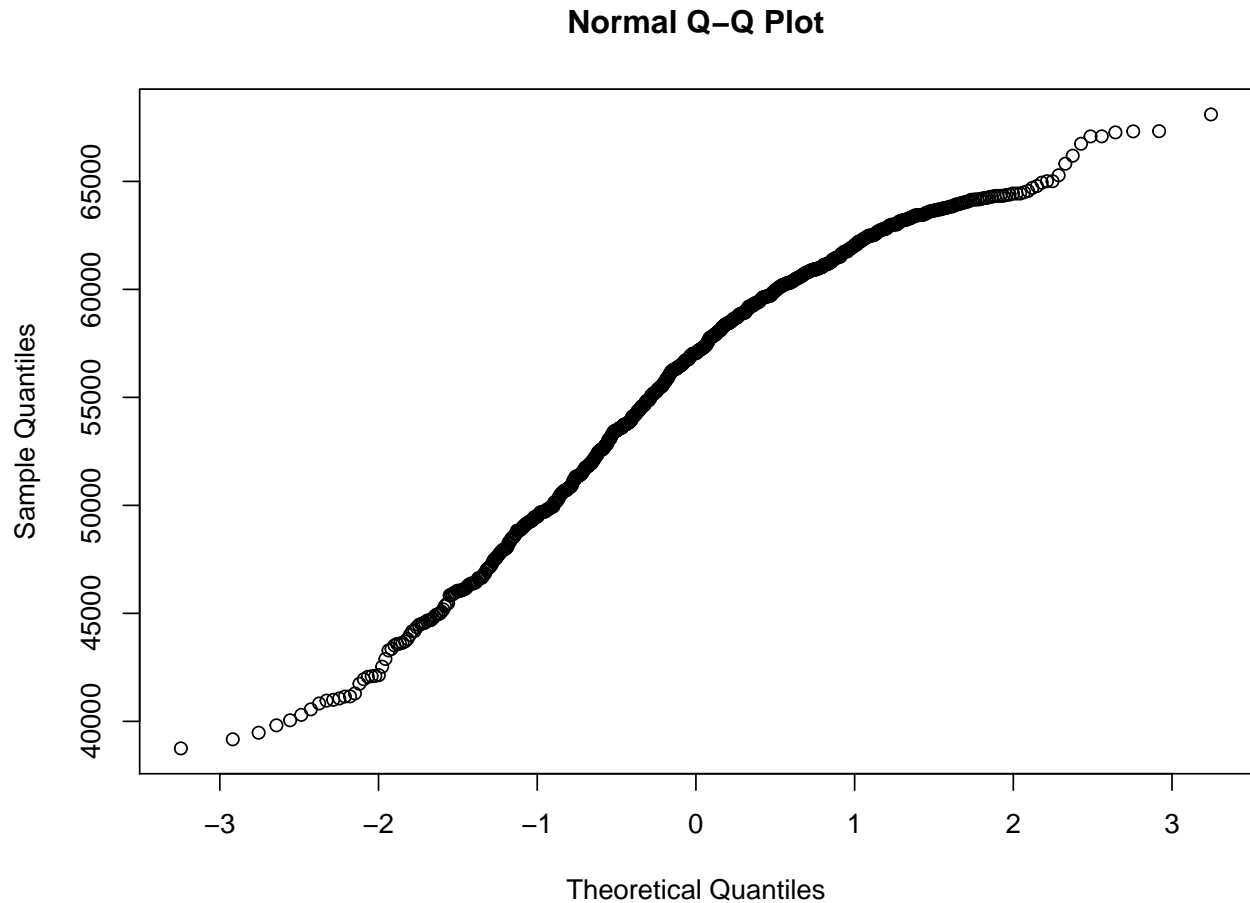
Normal Q–Q Plot


```
##              norm.test        pvals
## 1      Shapiro-Wilk:  4.542474e-14
## 2 Anderson-Darling:  1.531689e-23
```

Both normality tests generate incredibly small p values, rejecting normality. The Q-Q plot is very curved, where a normal one will follow a straight line. Therefore non-parametric tests are required to test for statistical significance.

The non-parametric Mann-Whitney U Test (wilcox.test) is used and result in very small p-values to demonstrate a statistically significant different in male/female times:

```
## [1] "wilcox.test p-value:" "0.000150214031214457"
```

Because the p value is incredibly small, it is clear that there is a significant difference between male and female finishing time.

Plot of male versus female ages
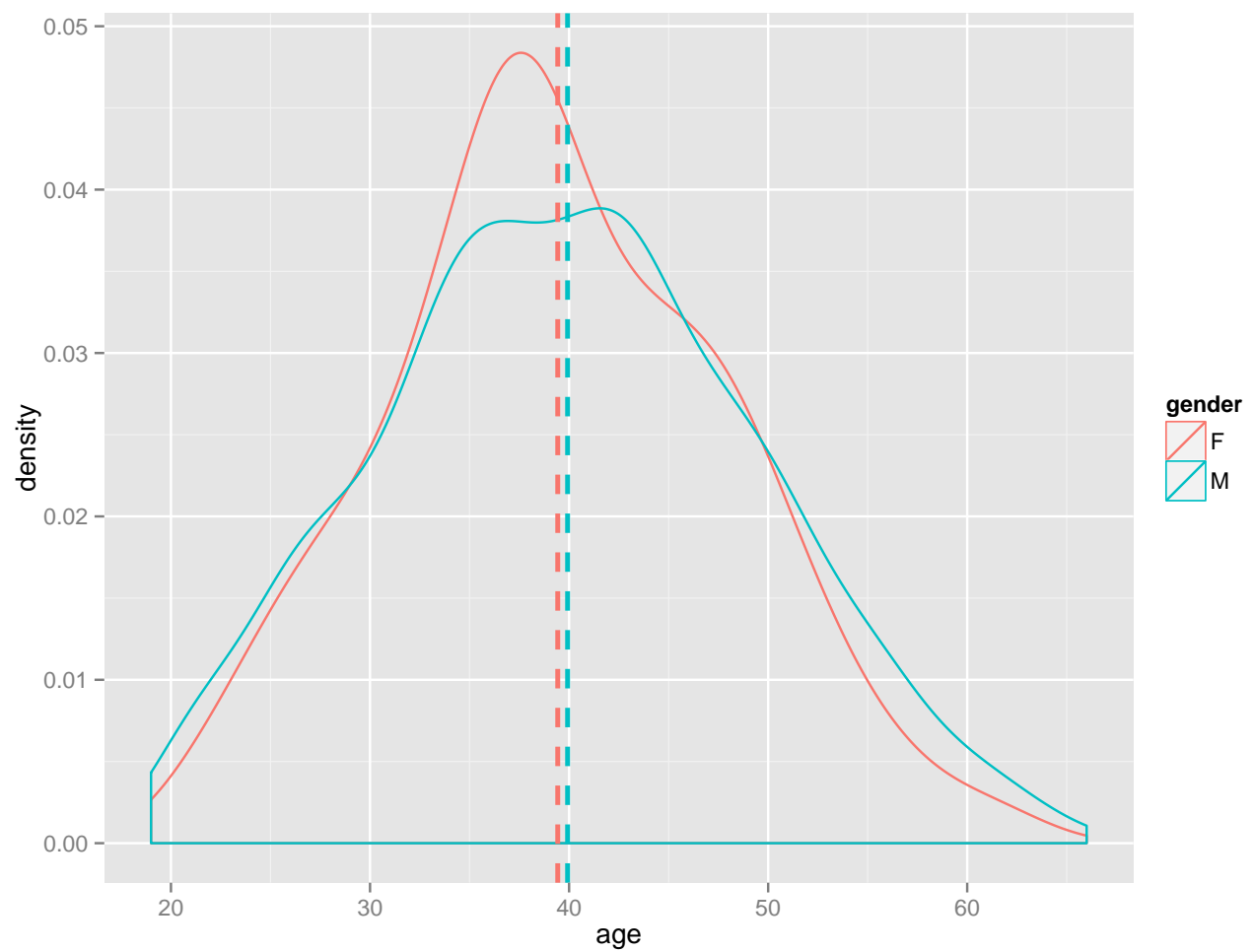
The age distributions look identical, and normally distributed. Running a check on the normality for the Female ages yields the following:
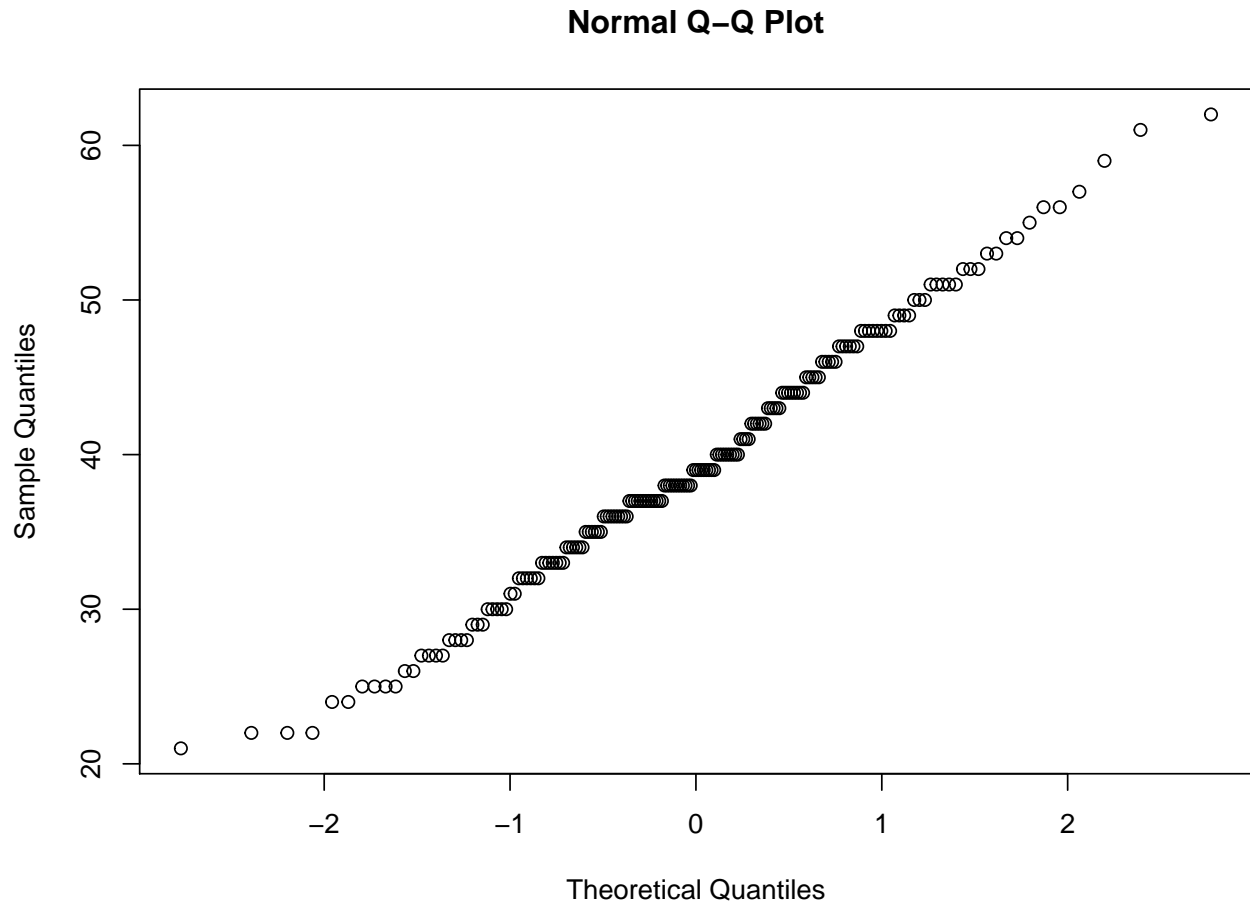
Normal Q–Q Plot


```
##            norm.test      pvals
## 1     Shapiro-Wilk:  0.3903384
## 2 Anderson-Darling:  0.3663799
```

The p values indicate normality, and the Q-Q plot confirms this. Therefore the Students T-test will be used. Based on the p value obtained below (0.49), there is no statistically significant difference between the ages within each group.

```
## [1] "t.test p-value:"    "0.487897806045054"
```

Age versus Hellgate Finishing Time

Looking at overall age, older runners tend to take longer (on average) to finish. Obviously this is not a surprising result, but from the scatter of the data, the trend is very gradual, meaning that older runners are still very competitive.
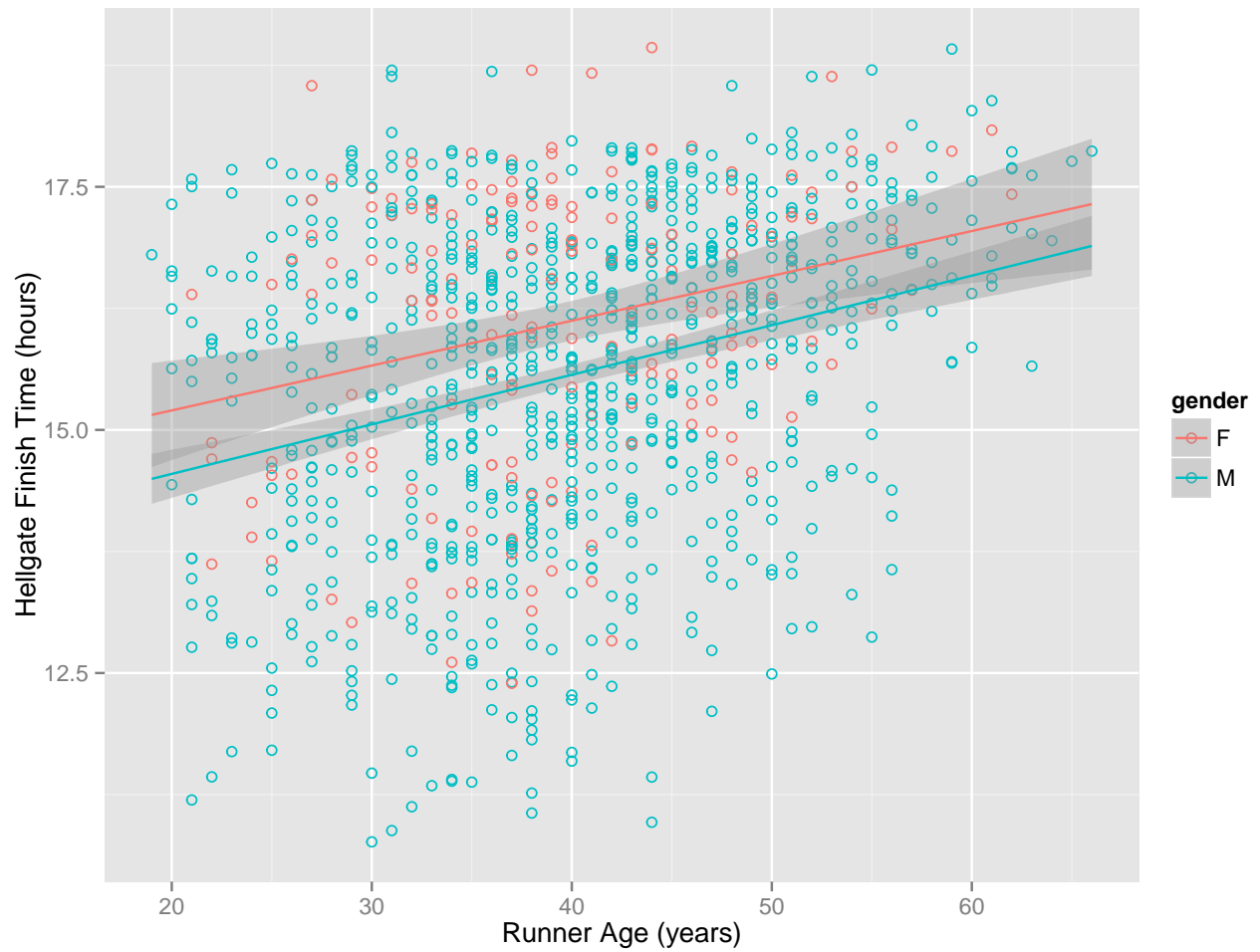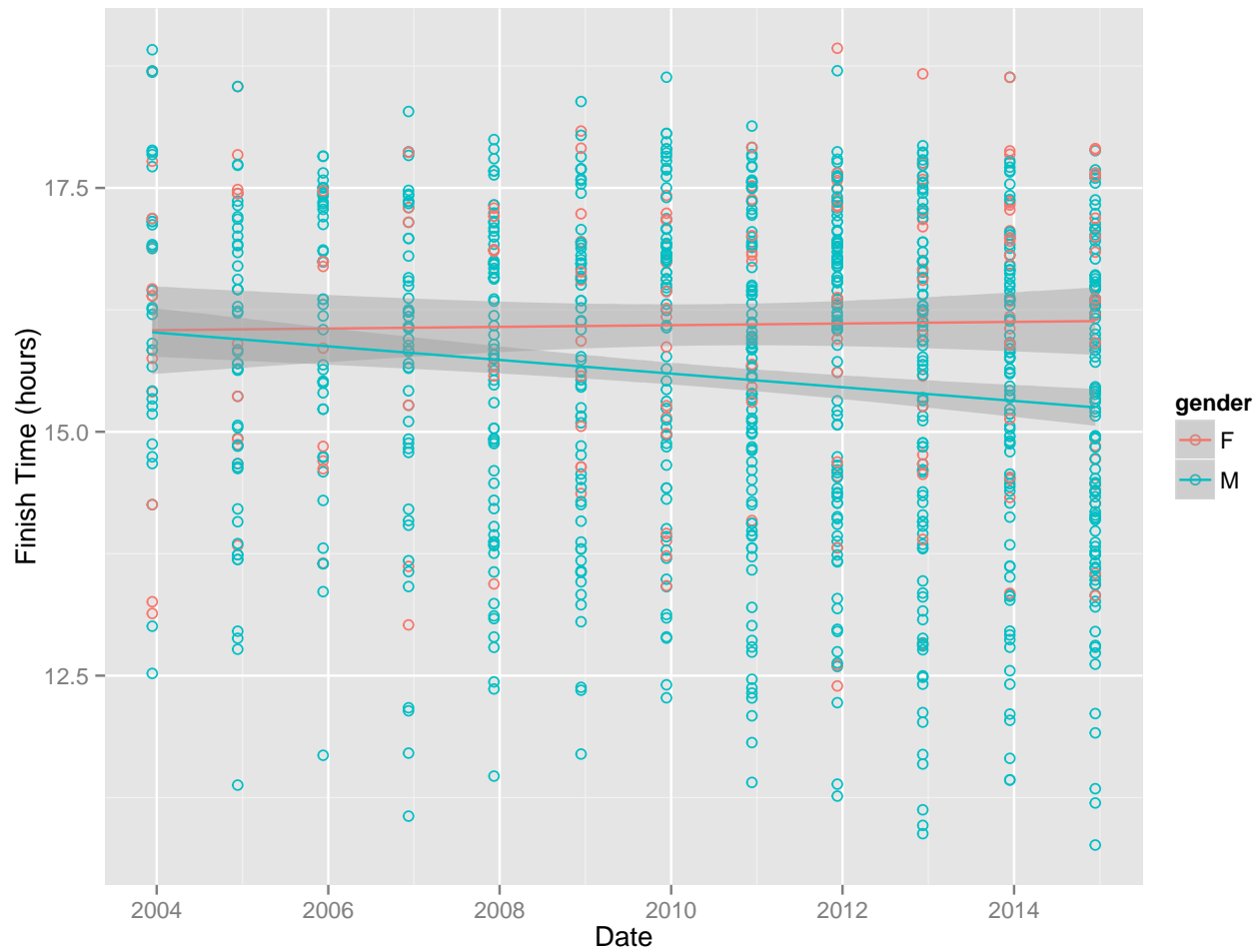
Age by race year, males vs females

In general, for both genders, the trend is an older. There are many veterans that have been running this race for several years, meaning that a core group of steadily older persons may be driving the numbers. However, the change is not that great.
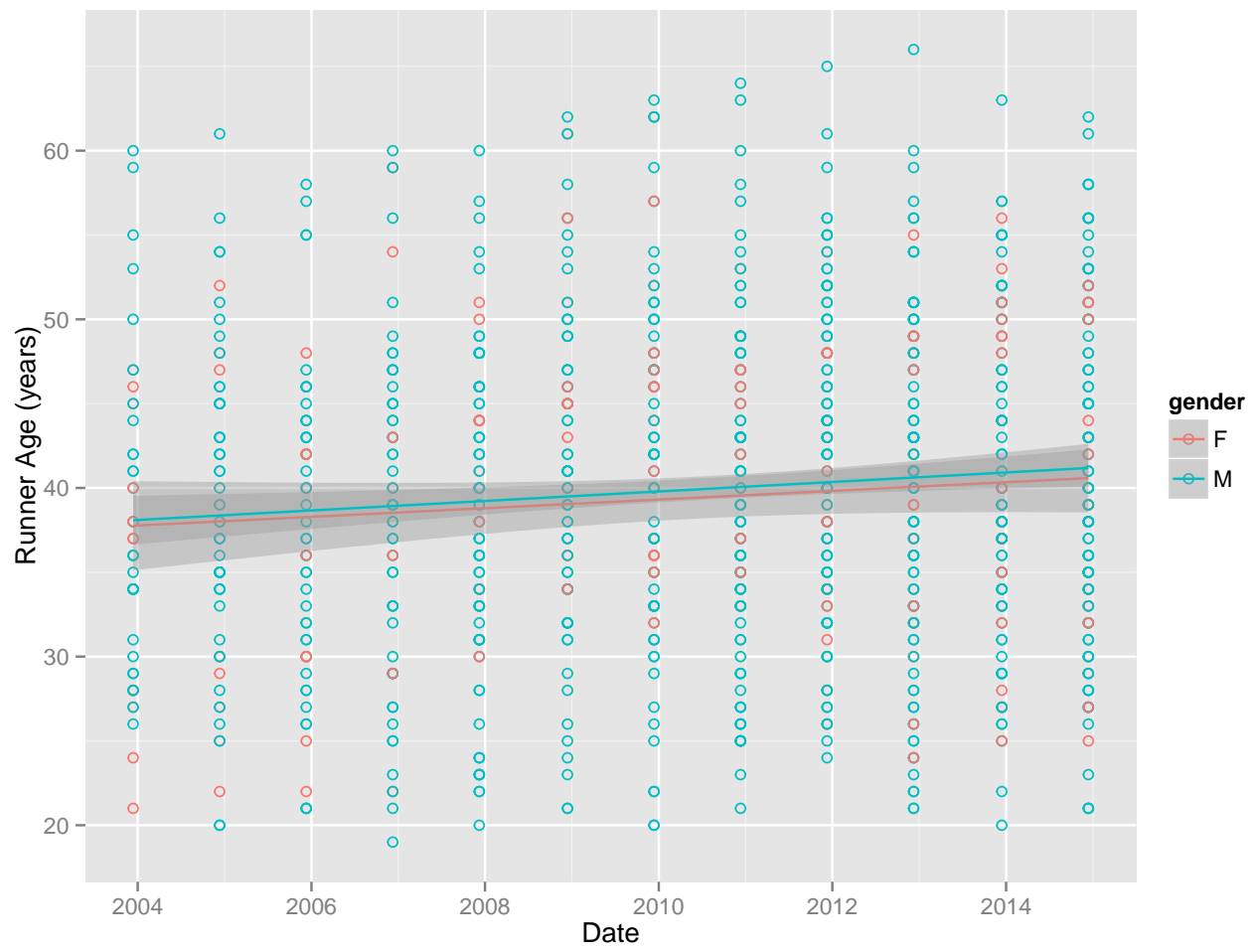
Times of men and women versus Hellgate finish time

In the 2008 race, the standard error shown by the gray shaded area (95% confidence interval around the mean) for the men's mean finishing time is distinctly different than the women's mean time (plus standard error). Overall, the women's mean finishing time seems to hold flat or increase slightly. The men's time continues this trend for every race after. First compare the male/female times before 2008:
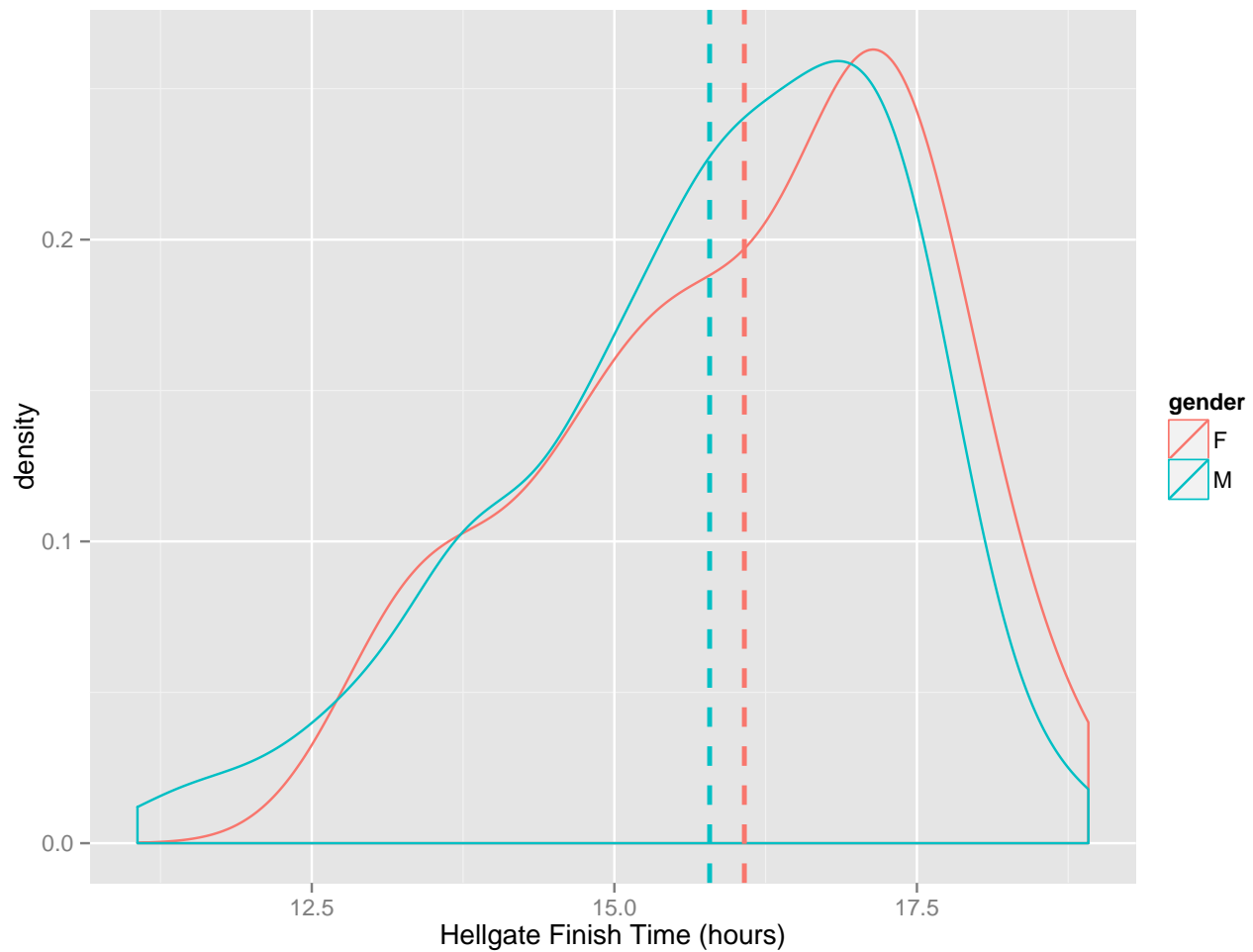
Testing pre and post 2008 Male/Female time differences:

```
##              group males.vs.females
## 1 2003-2007 p-value:       0.2428249975
## 2 2008-2014 p-value:       0.0001558289
```

Based on the p values obtained, for the first five years of the race (2003 - 2007) there is no significant difference between the male and female finishing times. After 2007, it's a different story, and the p-values are incredibly low, meaning that the males are significantly faster than females from 2008-2014.

Effect of Temperature on Finishing Percentage

The temperature with the calculated wind chill was used to understand the apparent effects of temperature experienced by the runners and overall finish rate. It appears that there might be a correlation between temperature and finishing percentage as seen in this graph:

Prediction Equation : $y = 65.7 + 0.525x,\ r^2 = 0.31$
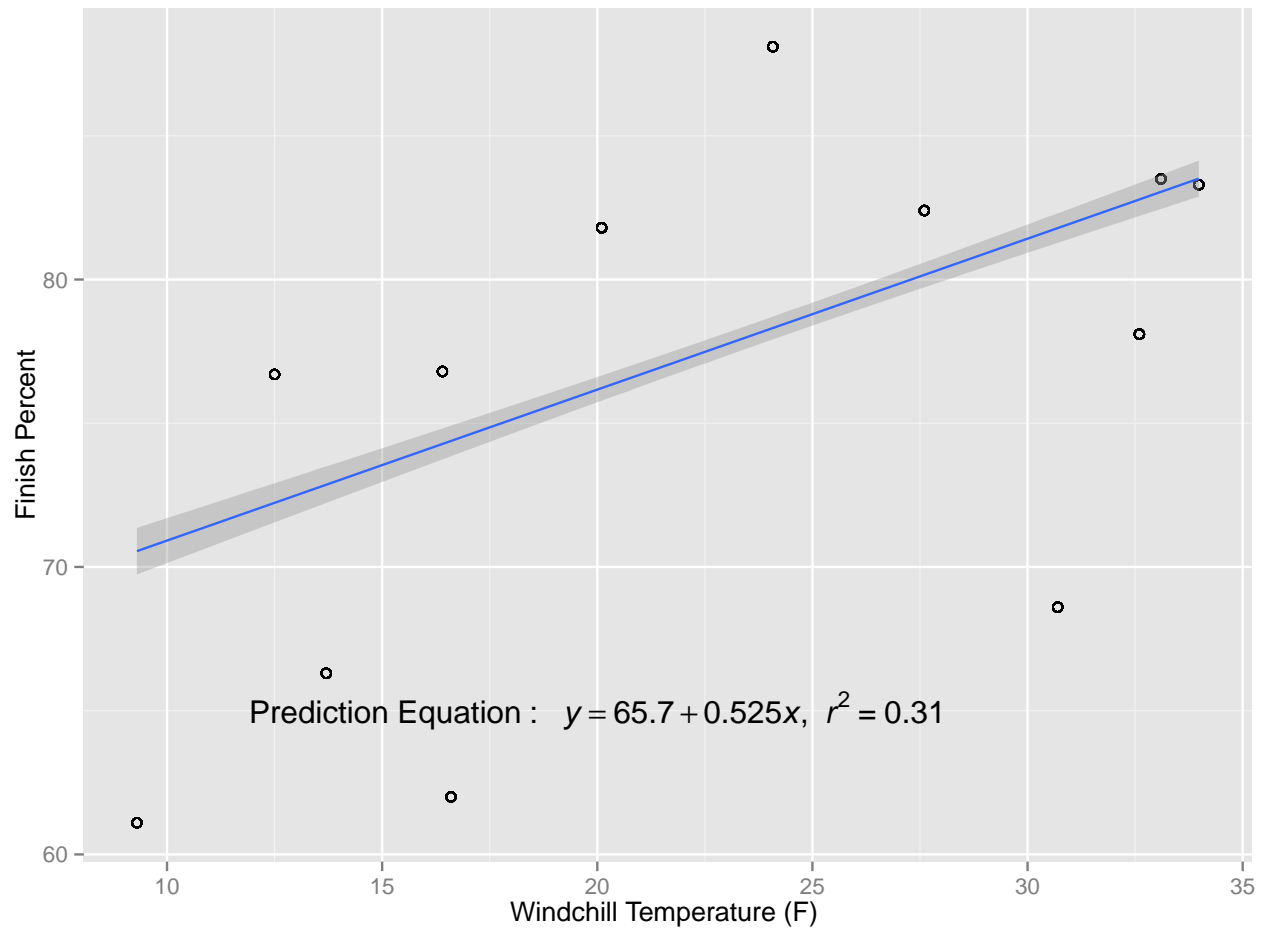
It would appear that temperature might influence the finishing percent, albeit the R^2 value is a bit low. Looking at the residual plots:
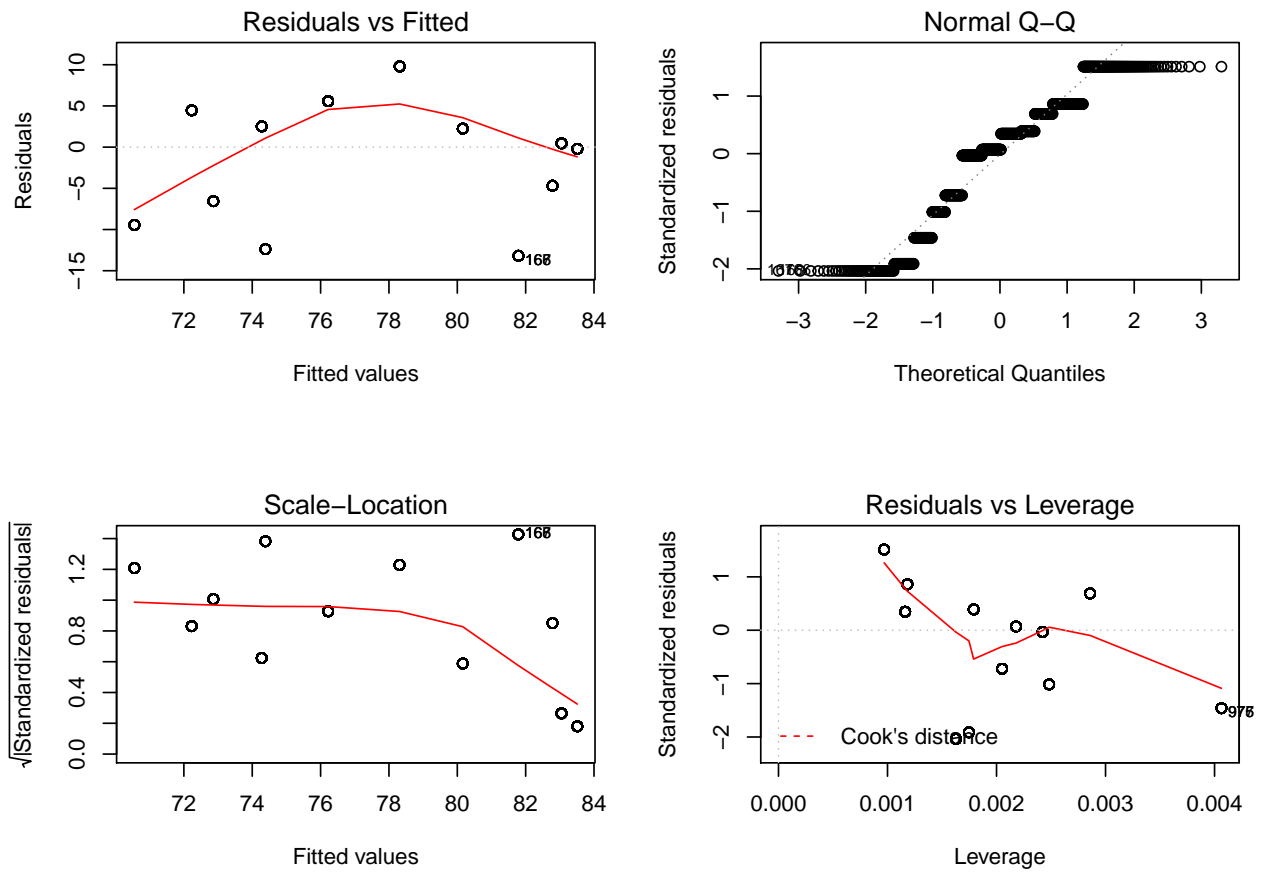
The Q-Q plot shows the data is not normal and the Residual plot has a curved fit, meaning that there is non-linear behavior and this type of obvious pattern means it is a poor predictor. Therefore, it is determined that temperature is not a good predictor on the finishing rate.

Effect of Course Conditions on Finishing Time

There are many factors that can influence the finishing times, including the overall course conditions.

Looking at a box-plot of times by course conditions, the average finishing times do not look that different, although the spread of the times appears to get more narrowly focused:
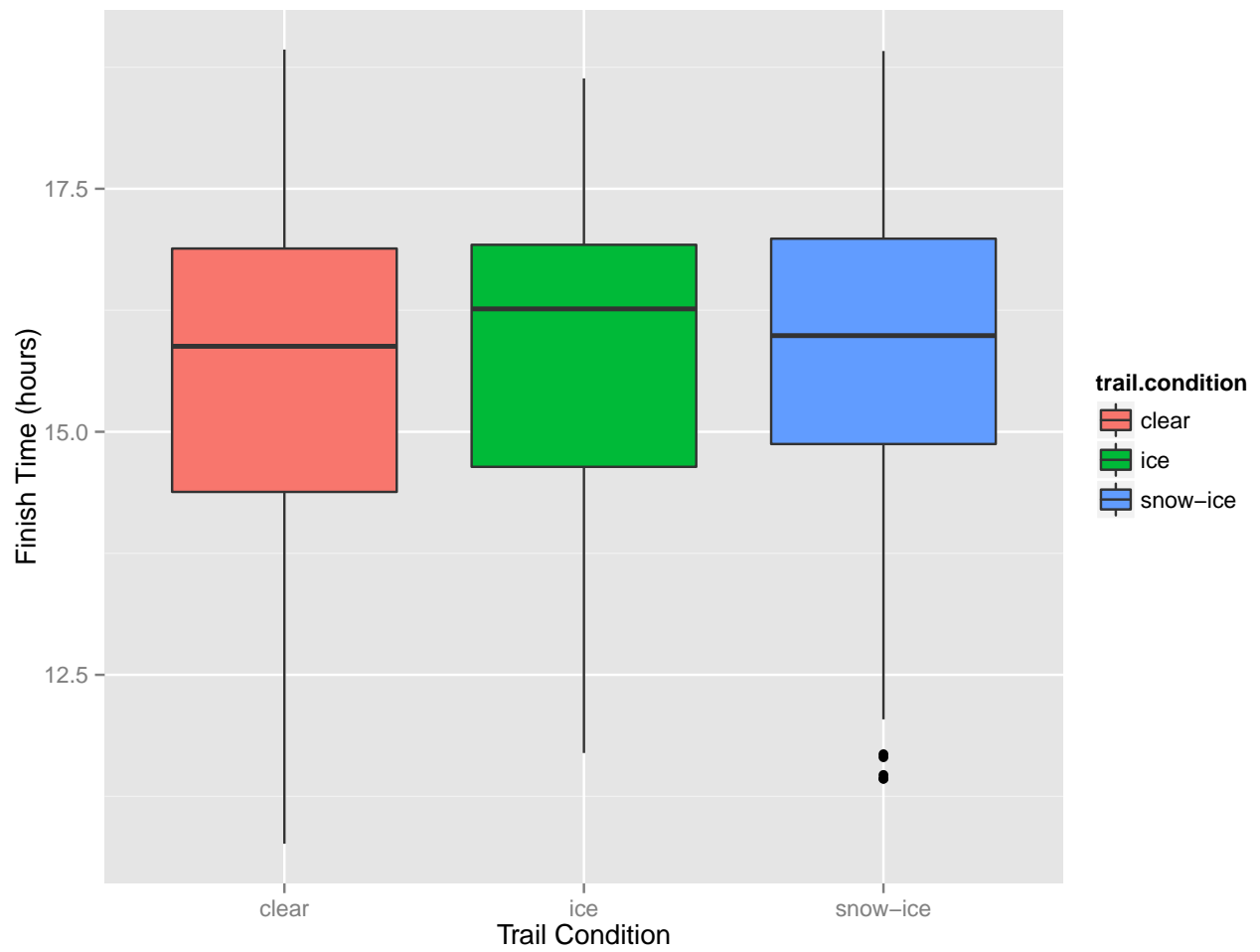
Looking at at histograms of each, they all look stacked on top of each other:
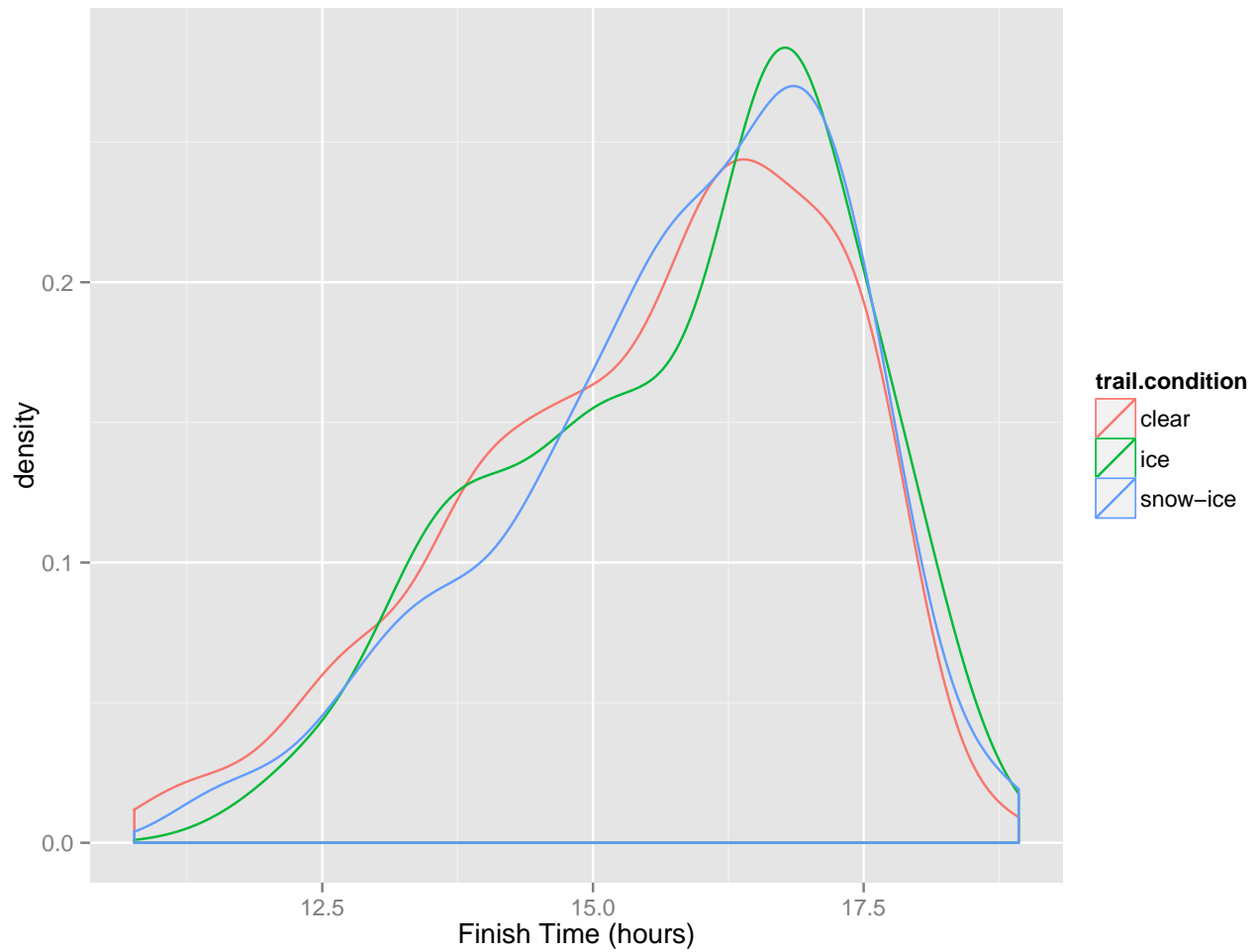
The data appears to be non-normal. The "Clear" trail condition looks the most "rounded" like a bell curve, the others are more skewed to the right. Just to test this hypothesis, three methods are provided for the "Clear" data that demonstrate non-normality. The Shapiro-Wilk test and the Anderson-Darling tests combined with a Q-Q plot provides the results:
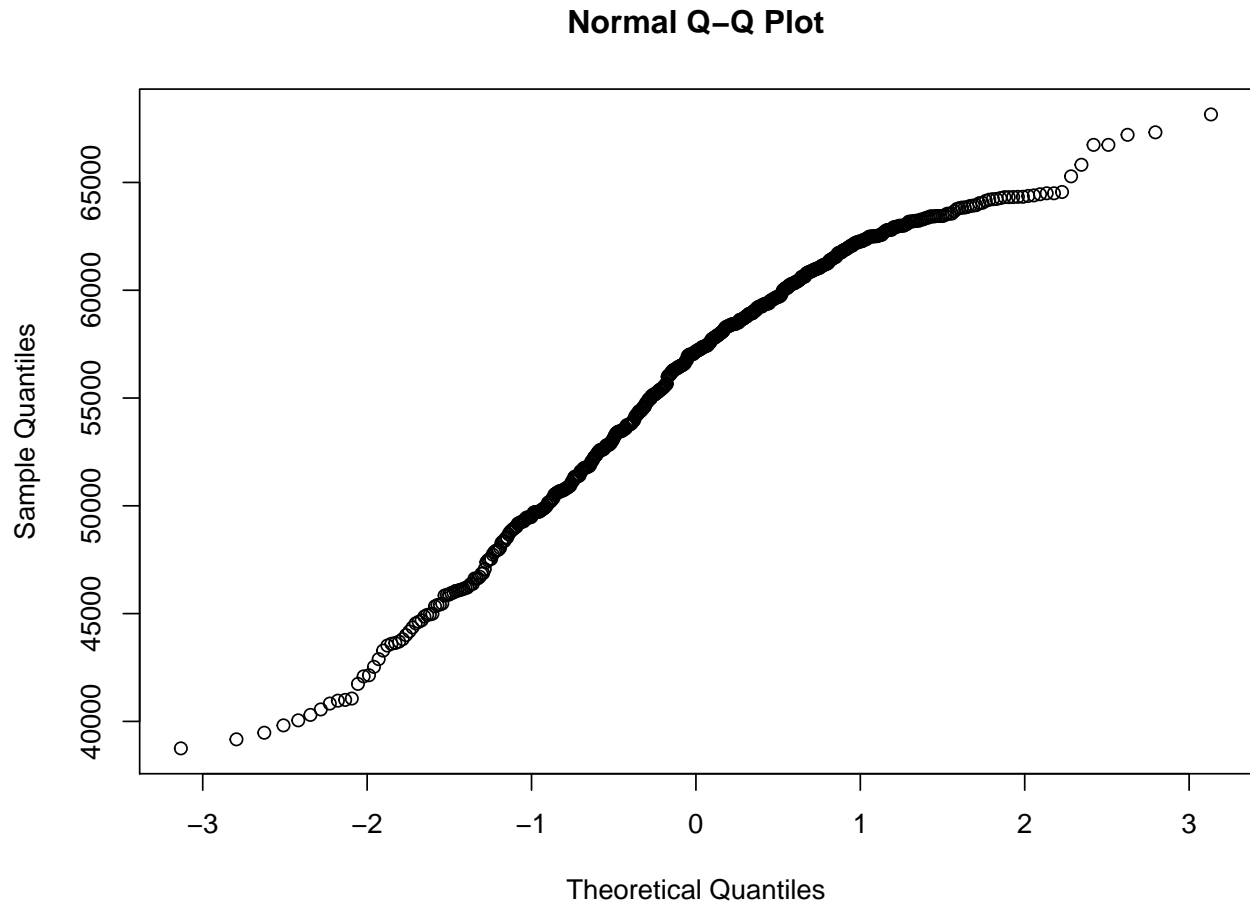
Normal Q–Q Plot


```
##              norm.test        pvals
## 1      Shapiro-Wilk:  3.782605e-11
## 2 Anderson-Darling:  8.475425e-16
```

Because both normality tests combined with the Q-Q plot showing very curved behavior is observed, the data is considered to be non-normal and a non-parametric test must be used.

Applying the Mann-Whitney U Test (a non-parametric test) to determine if any significant differences exist:

```
##              comparison       pvals
## 1      Clear vs. Ice:  0.07652347
## 2 Clear vs. Snow-Ice:  0.06402921
## 3   Ice vs. Snow-Ice:  0.85804870
```

As seen from the p-values produced, none are significant ($p<0.05$), and therefore the conclusion is that trail conditions are a factor in influencing finish time is rejected.

Beast Series Analysis

Because runners were faster overall after 2008, investigations into the Beast Series as a potential factor. While running the Beast Series did not appear to be a factor in influencing an overall faster time at Hellgate for those runners, it was thought that previous performance may predict a runners finish at Hellgate.

Predicting Hellgate Time for the Beast Series runners

Using the previous Beast time, can we predict the Hellgate finish time?

Taking the Beast finishers, calculate their average time (pace) before, then see how well it compares to their Hellgate pace. 2013 was considered the "Mini-Beast" as the Grindstone 100 was cancelled due to the US Government shutdown that occurred. Total miles is normalized to pace/mile so the millage difference can be accounted for. Mileage is done with the "advertised" race distance, meaning that Hellgate is 62 miles, Grindstone is 100 miles, etc. The actual distances may be a little bit longer, but this difference does not affect the comparison, and was done for clarity or ease of use.

Prediction Equation : $y = 7.43 + 0.555x$, $r^2 = 0.59$
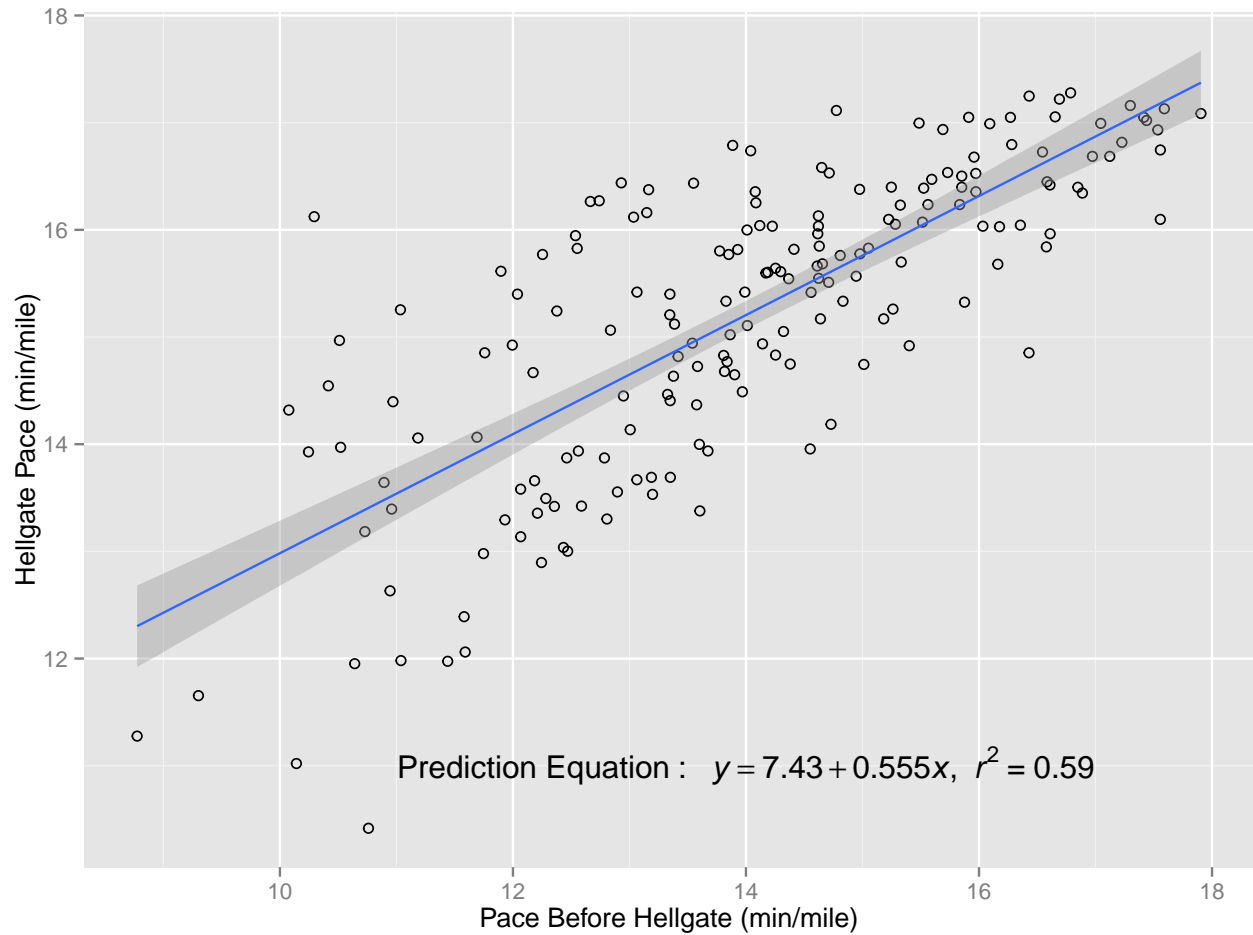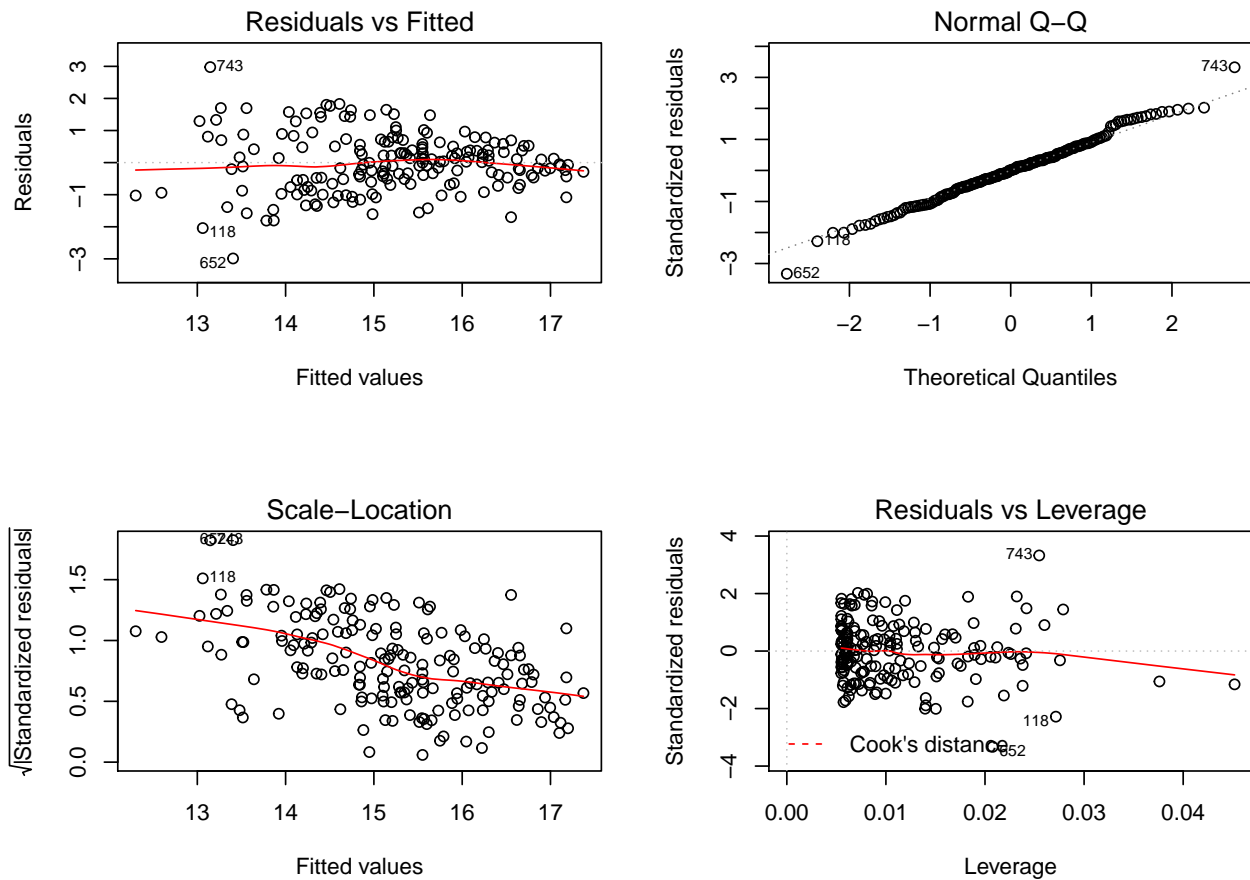
At first glance, this looks like a pretty good fit, but the R^2 value would indicate that the pave before Hellgate explains about 58% of the variation seen. The residual and Q-Q plots also look decent.

But, in the upper left, there appears to be a cluster of times that are much higher than what would be predicted. By looking at many different factors, including gender, temperature, number of times finishing Hellgate, and trail conditions, one thing stood out. The "snow-ice" year (2013) seems to explain that cluster as seen by the graph below:
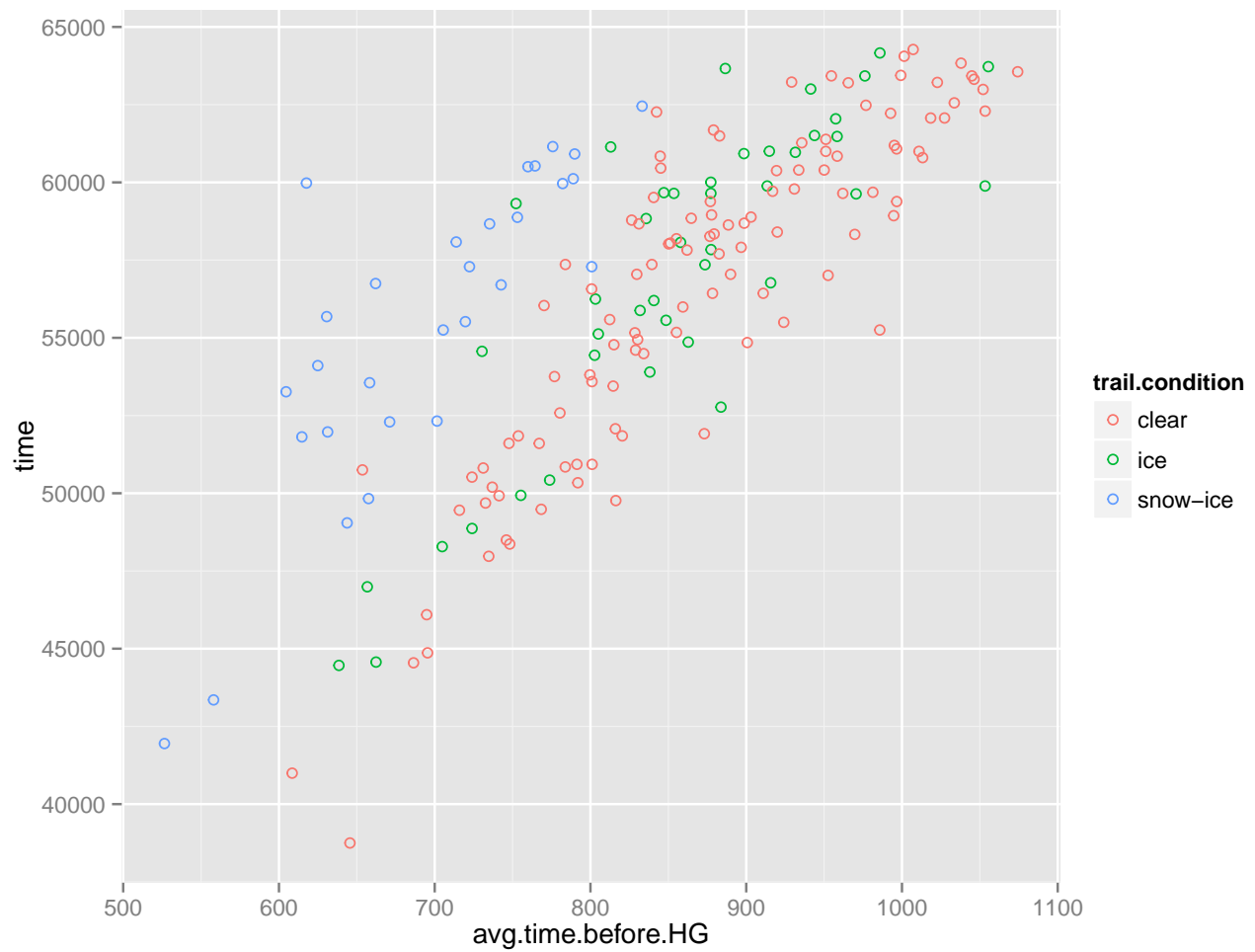
2013 certainly seems like a distinctly different year, as all of the finishing times are clustered above the others as seen in the above graph.

Another factor is experience, not just running Hellgate but doing what it takes to finish the Beast Series. Looking at the veteran Beast series runners, by Beast Series Finishes:
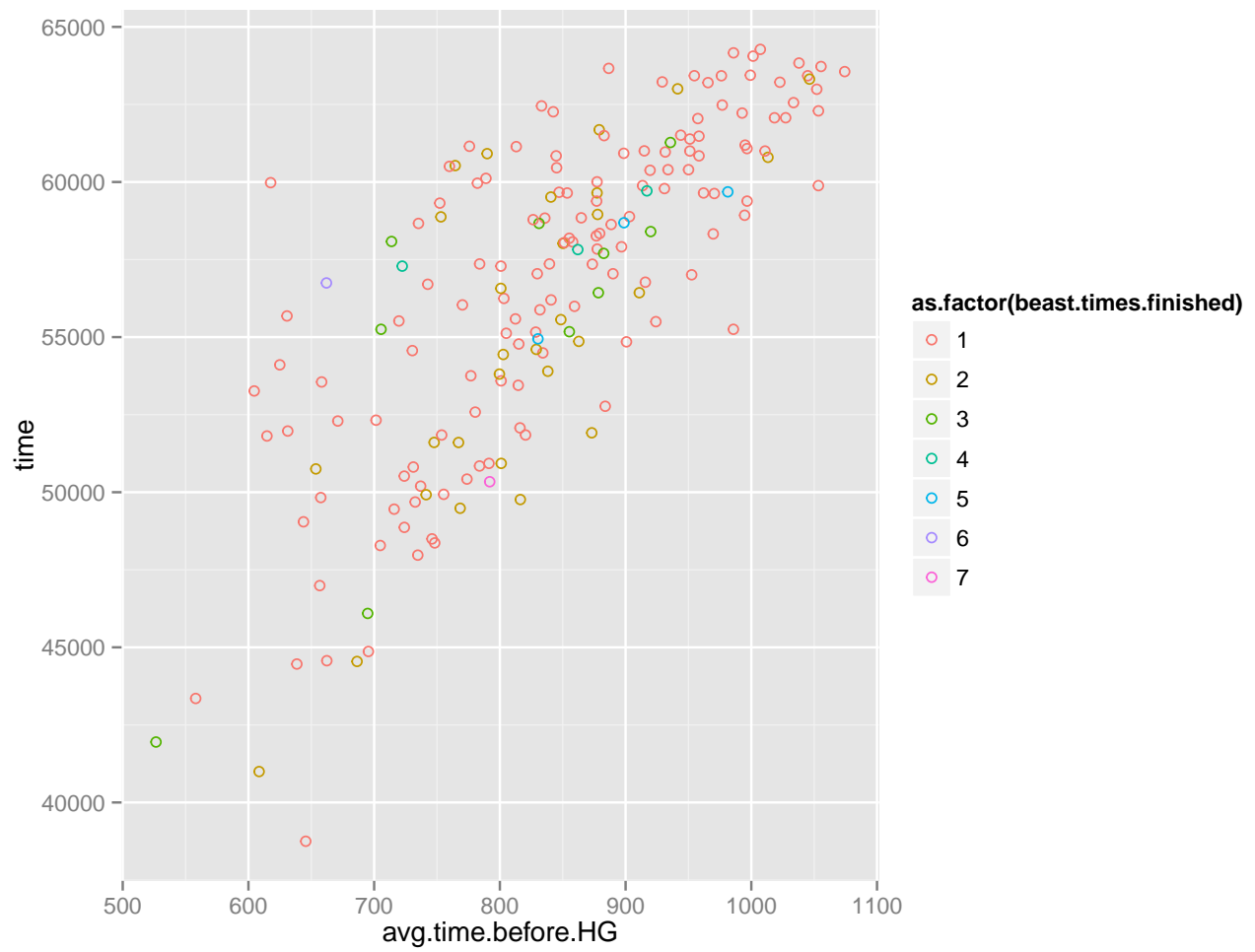
Now let's highlight Hellgate Finishes instead, showing that veterans of Hellgate are scattered throughout:
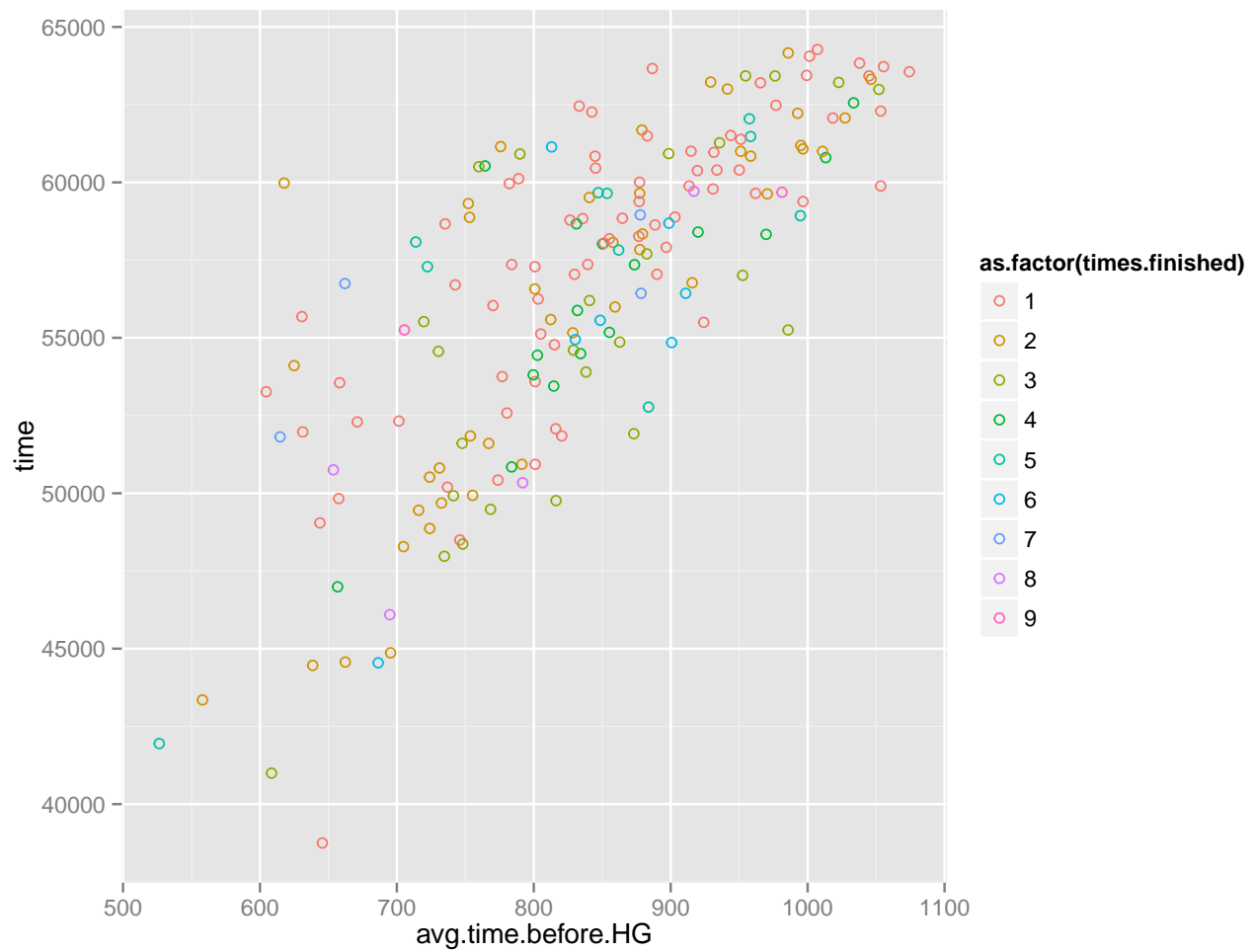
No distinct pattern from the number of Beast or Hellgate finishes is apparent.

2013 was also the year of the Govt shutdown, which forced Grindstone to be canceled. If years with and without Grindstone were treated separately, a better fit of the data is obtained:

No Grindstone : $y = 17701 + 54.3x$, $r^2 = 0.69$

With Grindstone : $y = 18229 + 44.4x$, $r^2 = 0.75$
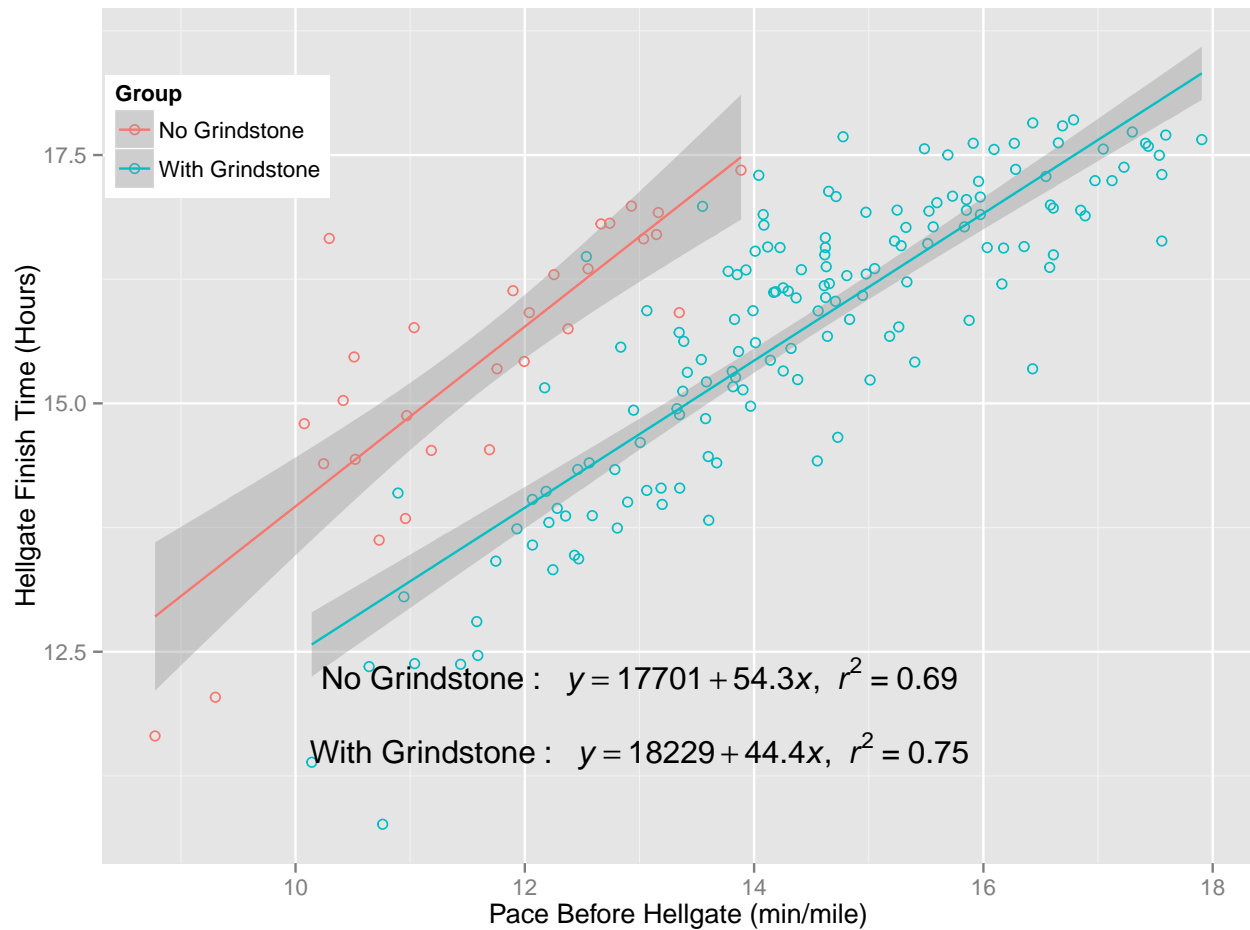
The prediction equation is significantly improved for each group. Training is important, and race experience on a tough course would have been good preparation. but a test to see if 2013 was statistically different needs to be done before removing the group from the sample.

Another question on whether or not it was the trail conditions or lack of a big race prior to Hellgate was to blame. There were four years which had snow-ice conditions. Just Looking at the years with snow-ice, the year in question (Y11, 2013) as the second fastest average time:
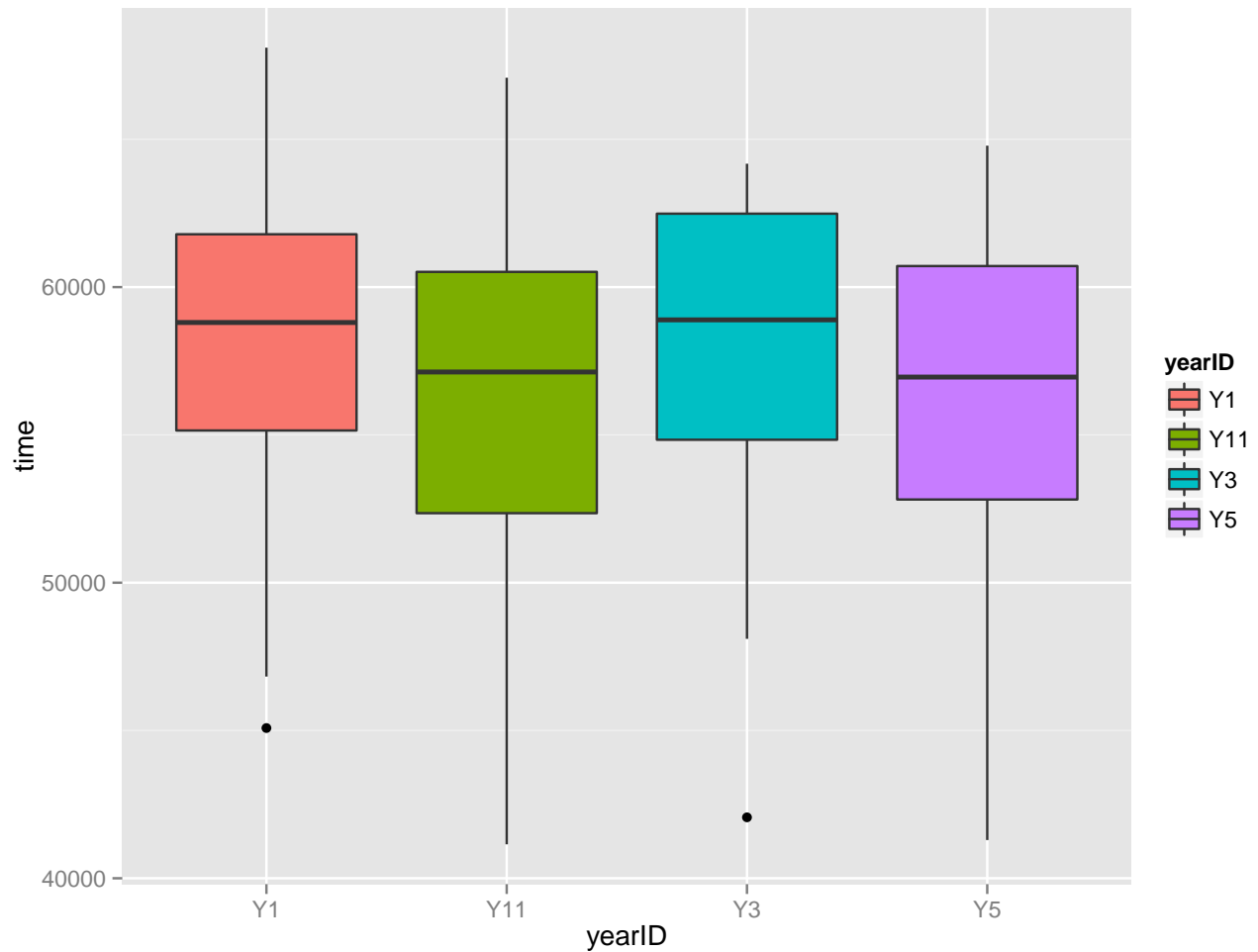
Several comparisons of 2013 can be made. First whether or not 2013 was different than other Beast years and from other Hellgate years. Next, looking at trail conditions between other snow-ice years and the other clear and ice years too. Again for non-normality, statistical significance is done with the Mann-Whitney U test:

```
##                 comparison.labels comparison.pvals
## 1        2013 vs other Beast years       0.17660283
## 2          2013 versus all Hellgate       0.60931047
## 3             2013 vs. Other Snow-Ice       0.09194064
## 4 Snow-Ice vs. Clear and Ice Years       0.14950988
```

All of the p values are greater than 0.05, therefore there are no significant differences between the 2013 year and other Beast years, all other Hellgate years or between all other terrain conditions. Therefore, the 2013 group should not be treated as "outliers" in the data.

Comparing Beast Runners with other Hellgate Runners

While the 2013 group may not be statistically different, looking at all Beast years is important to understand if any differences can be observed. A box-plot of the years with the Beast Series Runners compared to the rest of the Hellgate runners is below:
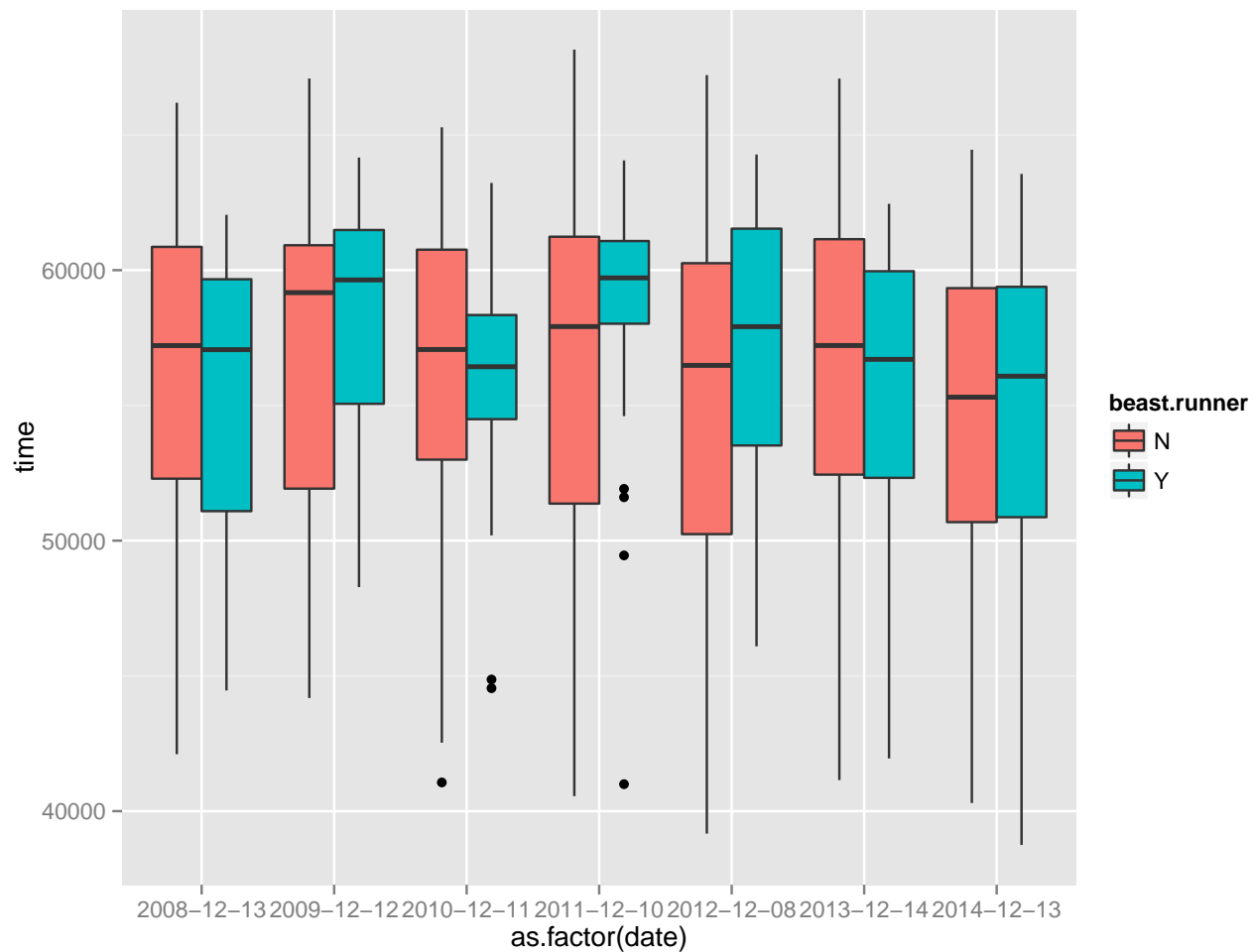
For each year, testing to see if there are any differences between Beast Runners and the rest of the runners:

```
##   race.year      pvals
## 1      2008 0.4306507
## 2      2009 0.4346617
## 3      2010 0.8212902
## 4      2011 0.1086852
## 5      2012 0.2282386
## 6      2013 0.3269300
## 7      2014 0.6129500
```

In all years, the p values are greater than 0.05, meaning that there are no significant differences between Beast Runners and other Hellgate Runners.

However, in 2011, the 9th year, the box-plot shows that the beast runners might have been significantly slower than their peers. There are some outliers in this group. A histogram of this year is below and a statistical test would claim that the difference is not significant. Removing the outliers and re-running the test make this year significantly different, that the average Beast Runner was slower, on average, from the rest of the pack.
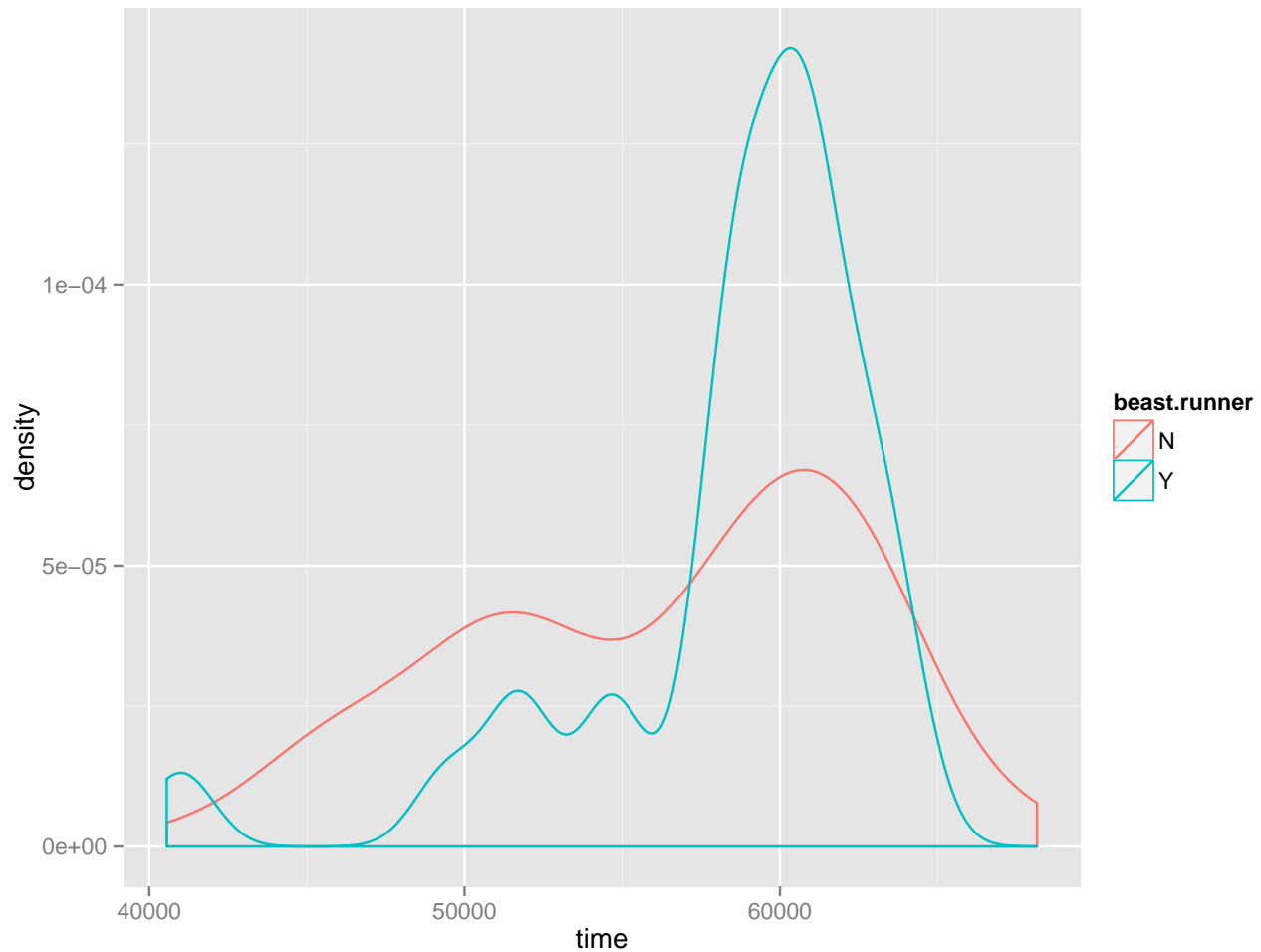
```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  dfHG_all$time[which(dfHG_all$yearID == "Y9" & dfHG_all$beast.runner ==  and dfHG_all$time[whi
## W = 1445.5, p-value = 0.1087
## alternative hypothesis: true location shift is not equal to 0


##
##  Wilcoxon rank sum test with continuity correction
##
## data:  dfHG_all$time[which(dfHG_all$yearID == "Y9" & dfHG_all$beast.runner ==  and dfHG_all$time[whi
## W = 60, p-value = 0.0325
## alternative hypothesis: true location shift is not equal to 0
```

Developing a better model: Beast Series Analysis

Using the Total Beast time resulted in a 58% fit and the residual and Q-Q plots were decent, but it is an aggregate of all races. The three 50k races in the beginning of the year are mixed in and it may be possible to look more closely at each race leading up to Hellgate to determine if a better model can be produced.

The Beast data was obtained from both Eco-XSports and Extreme Ultra Running, depending on the year (early years can only be found on Extreme Ultra Running). First, a linear model was fit using a combination of all the race finishing times. Next, using the "step" function in R, the software tests several combinations

to determine the outcomes that are most significant and optimizes the model with the best parameters to form a prediction equation.

```
## Start:  AIC=2422.05
## beast.Hellgate ~ beast.HL.50K + beast.T.MTN + beast.PL.50K +
##     beast.GS.100 + beast.MMTR
##
##                Df Sum of Sq         RSS    AIC
## - beast.T.MTN   1     697546  877796833 2420.2
## - beast.HL.50K  1    1127697  878226984 2420.2
## - beast.PL.50K  1   10429774  887529061 2421.9
## <none>                        877099287 2422.1
## - beast.GS.100  1  204355624 1081454911 2452.5
## - beast.MMTR    1  379745149 1256844436 2475.8
##
## Step:  AIC=2420.17
## beast.Hellgate ~ beast.HL.50K + beast.PL.50K + beast.GS.100 +
##     beast.MMTR
##
##                Df Sum of Sq         RSS    AIC
## - beast.HL.50K  1     654894  878451727 2418.3
## <none>                        877796833 2420.2
## - beast.PL.50K  1   12763825  890560658 2420.4
## + beast.T.MTN   1     697546  877099287 2422.1
## - beast.GS.100  1  204971861 1082768694 2450.7
## - beast.MMTR    1  401378178 1279175011 2476.5
##
## Step:  AIC=2418.29
## beast.Hellgate ~ beast.PL.50K + beast.GS.100 + beast.MMTR
##
##                Df Sum of Sq         RSS    AIC
## <none>                        878451727 2418.3
## - beast.PL.50K  1   14593204  893044930 2418.8
## + beast.HL.50K  1     654894  877796833 2420.2
## + beast.T.MTN   1     224743  878226984 2420.2
## - beast.GS.100  1  206830343 1085282070 2449.1
## - beast.MMTR    1  414089376 1292541103 2476.2


##
## Call:
## lm(formula = beast.Hellgate ~ beast.PL.50K + beast.GS.100 + beast.MMTR,
##     data = beast.times)
##
## Residuals:
##     Min     1Q  Median      3Q     Max
## -6775.2 -1559.4  -269.9  1324.3  7048.9
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.259e+04  1.984e+03   6.347 2.44e-09 ***
## beast.PL.50K 1.765e-01  1.114e-01   1.584    0.115
## beast.GS.100 1.255e-01  2.104e-02   5.963 1.69e-08 ***
## beast.MMTR   7.050e-01  8.357e-02   8.437 2.45e-14 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2412 on 151 degrees of freedom
##   (29 observations deleted due to missingness)
## Multiple R-squared:  0.7952, Adjusted R-squared:  0.7911
## F-statistic: 195.4 on 3 and 151 DF,  p-value: < 2.2e-16
```

Residuals vs Fitted

Normal Q–Q

Scale–Location

Residuals vs Leverage
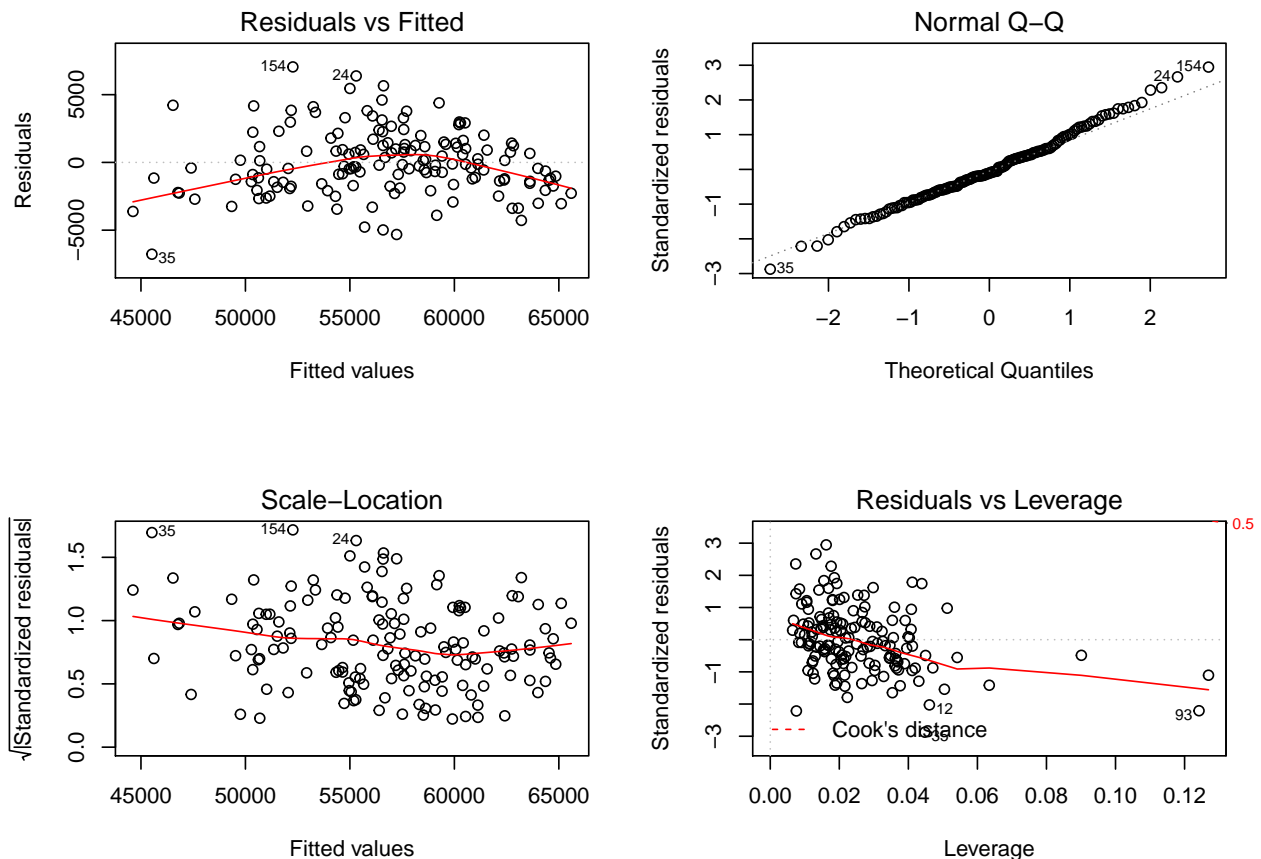
The output indicates that the Grindstone and MMTR time are significant, but the addition of the Promise Land (PL.50K) factor seems weird. This was suspect to be "over fitting" and a new model with only the interaction of the Grindstone and MMTR results would be used.

Looking at the interaction of MMTR and GS100 to predict Hellgate Times:

To begin, summaries of the linear fits to compare MMTR and Grindstone alone, as well as their interaction:

```
##
## Call:
## lm(formula = Hellgate ~ MMTR, data = beast)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10978.0 -1999.0   -56.4  1876.4  7869.9
## attr(,"class")
## [1] "Duration"
## attr(,"class")attr(,"package")
```

```
## [1] "lubridate"
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.258e+04  2.077e+03   6.055 7.84e-09 ***
## MMTR        1.174e+00  5.517e-02  21.284  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2817 on 182 degrees of freedom
## Multiple R-squared:  0.7134, Adjusted R-squared:  0.7118
## F-statistic:    453 on 1 and 182 DF,  p-value: < 2.2e-16


##
## Call:
## lm(formula = Hellgate ~ GS.100, data = beast)
##
## Residuals:
##      Min      1Q  Median      3Q     Max
## -9567.6 -2253.6   157.6  2222.9  7578.9
## attr(,"class")
## [1] "Duration"
## attr(,"class")attr(,"package")
## [1] "lubridate"
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.719e+04  1.747e+03   15.57   <2e-16 ***
## GS.100      2.772e-01  1.622e-02   17.09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3104 on 153 degrees of freedom
##   (29 observations deleted due to missingness)
## Multiple R-squared:  0.6563, Adjusted R-squared:  0.654
## F-statistic: 292.1 on 1 and 153 DF,  p-value: < 2.2e-16


##
## Call:
## lm(formula = Hellgate ~ MMTR * GS.100, data = beast)
##
## Residuals:
##      Min      1Q  Median      3Q     Max
## -6855.3 -1308.4  -122.6  1224.7  6910.4
## attr(,"class")
## [1] "Duration"
## attr(,"class")attr(,"package")
## [1] "lubridate"
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.948e+04  1.096e+04  -2.689 0.007970 **
## MMTR         1.897e+00  2.973e-01   6.381 2.04e-09 ***
## GS.100       5.684e-01  1.097e-01   5.180 6.99e-07 ***
```

```
## MMTR:GS.100 -1.128e-05  2.859e-06  -3.946 0.000121 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2315 on 151 degrees of freedom
##   (29 observations deleted due to missingness)
## Multiple R-squared:  0.8112, Adjusted R-squared:  0.8075
## F-statistic: 216.3 on 3 and 151 DF,  p-value: < 2.2e-16
```

Note that the model excludes the missing GS100 values from the 2013 year when GS100 was cancelled due to the Govt Shutdown.

To Summarize the (adjusted) R^2 values: 1) MMTR R^2 = .71 2) GS100 R^2 = .654 3) Interaction of MMTR and GS100 R^2 = .08075

This means that the Hellgate data can be plotted against a new set of times from the combination of MMTR and GS100 based on the formula below:

Hellgate time (predicted) = (-2.948e+04) + (1.897)MMTR.time + (5.684e-01)GS100.time + (-1.128e-05)MMTR.time$GS100.time

Note that the coefficients from the interaction model were used.

Using the interaction equation, generating a plot and new model:
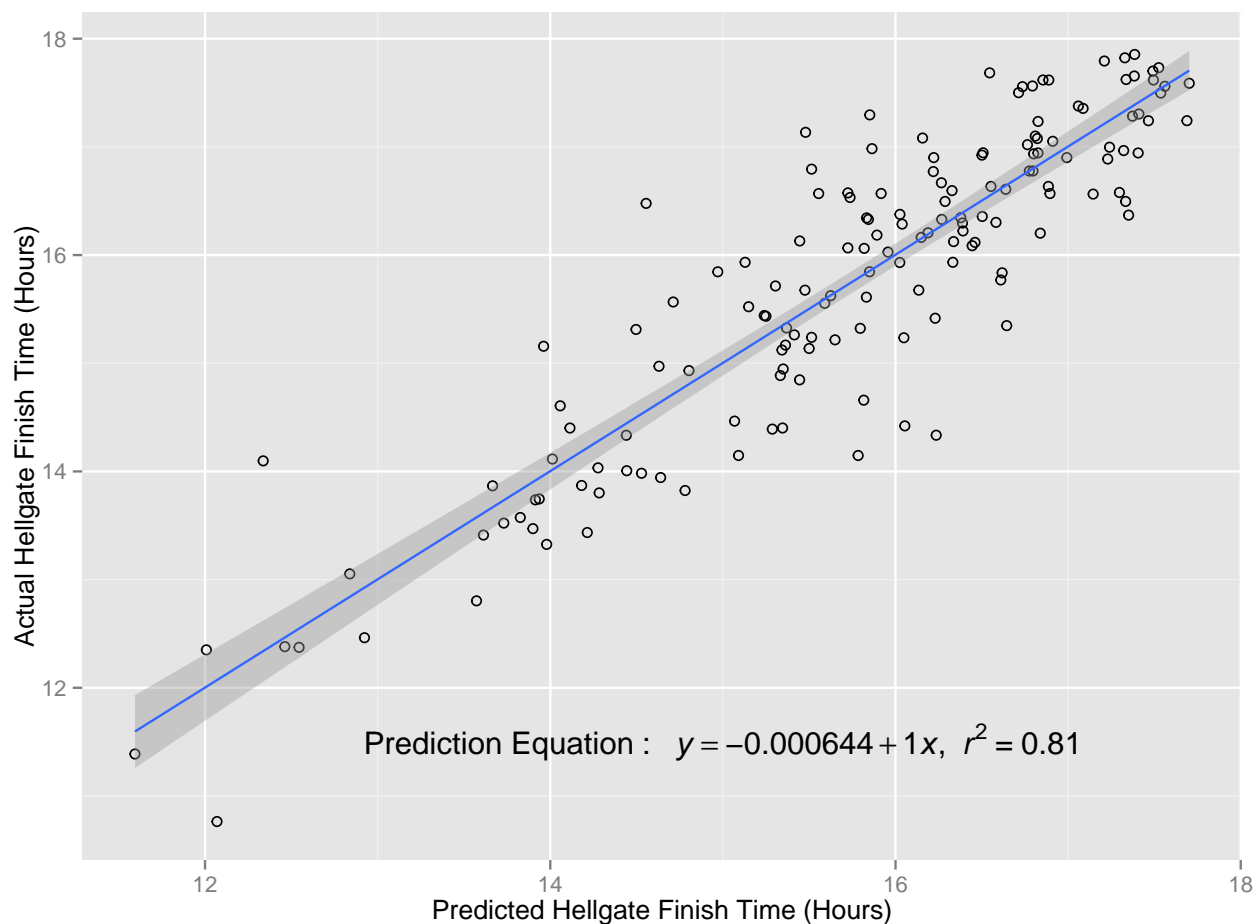
Looking at the residual plots of the model to see if there are any biases or other abnormalities. While the X axis on the residual plot looks slightly unbalanced, it has no other obvious pattern. The Q-Q plot has good normality, so the model looks good:
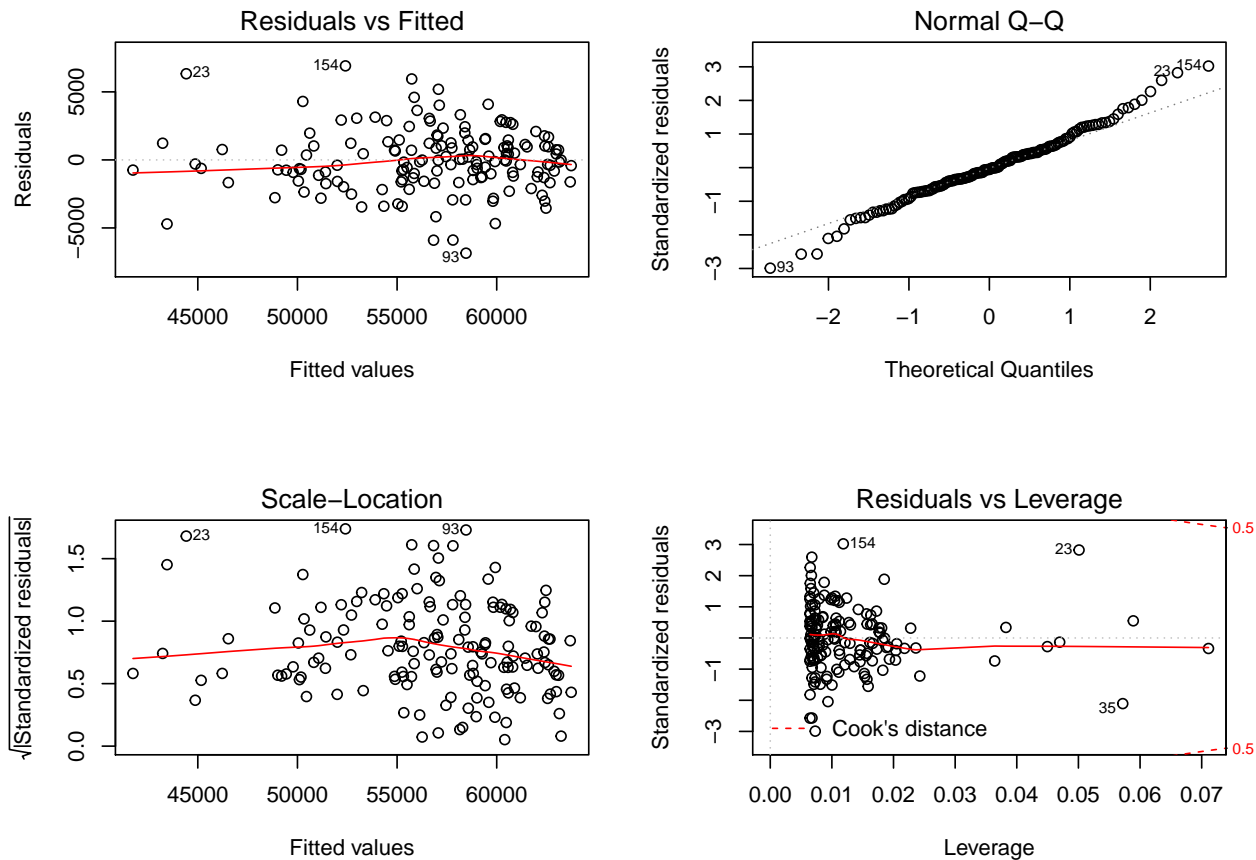
Based on this result, the results from Grindstone and MMTR can be used reasonable well to predict Hellgate performance.

References

1. cookbook-r (2015). Online resource. http://www.cookbook-r.com/Graphs/Plotting_distributions_ (ggplot2)

2. Eco-XSports (2015). Retrieved January, 2015 from www.eco-exsports.com

3. Extreme Ultra Running (2015). Retrieved January 2015 from www.extremeultrarunning.com

4. Jampen SC, Knechtle B, Rust CA, Lepers R, Rosemann T (2013). Increase in finishers and improvement of performance of masters runners in the Marathon des Sables. Int J Gen Med. 2013; 6: 427–438. http://dx.doi.org/10.2147/IJGM.S45265

5. Regression Equation (2015). Retrieved on January 18,2015 from: http://stackoverflow.com/questions/ 7549694/ggplot2-adding-regression-line-equation-and-r2-on-graph

6. Trim Whitespace (2015). Retrieved from http://stackoverflow.com/questions/2261079/how-to-trim-leading-and-trailing-w on January 27, 2015

7. Line Smoothing (2015). http://stats.stackexchange.com/questions/110380/smoother-lines-for-ggplot2. Retrieved January 28, 2015.

8. UltraRunning Magazine (2013). 2013 UltraRunning participation by the numbers. Retrieved on February 10, 2015 from the website http://www.ultrarunning.com/featured/2013-ultrarunning-participation-by-the-numbers/

9. Weather Data (2015). Menne, Matthew J., Imke Durre, Bryant Korzeniewski, Shelley McNeal, Kristy Thomas, Xungang Yin, Steven Anthony, Ron Ray, Russell S. Vose, Byron E.Gleason, and Tamara G. Houston (2012): Global Historical Climatology Network - Daily (GHCN-Daily), Version 3. Custom GHCN-Daily CSV: CITY:US510016 - Roanoke, VA US dates 12-1-2003 - 12-31-2014. NOAA National Climatic Data Center. doi:10.7289/V5D21VHZ Retrieved January 5, 2015.

10. Windchill Calculation (2015). Retrieved January 2015 from http://www.nws.noaa.gov/om/winter/windchill.shtml