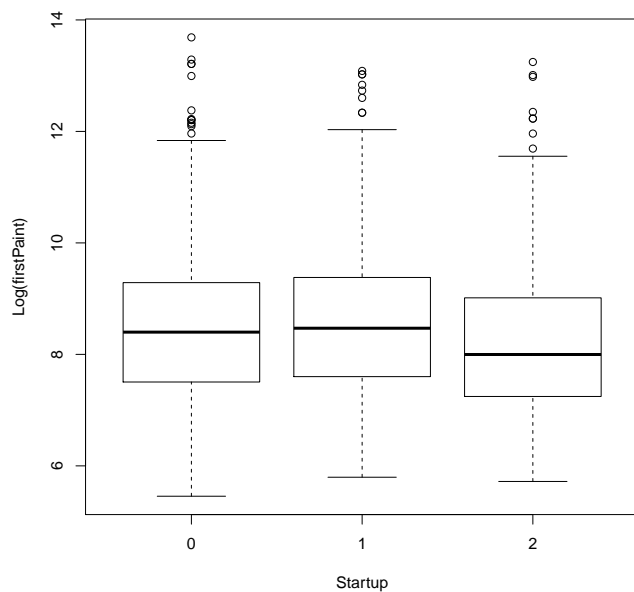This analysis is for the period 20120529 − 20120611, appName==Firefox, appUpdateChannel==nightly.

Descriptive Statistics

The boxplot and table show that observations with a startup (STARTUP_USING_PRELOAD) value of 2 have a lower firstPaint at each quartile.

Descriptive statistics on the original scale:

	startup	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	0	234	1817	4444	11580	10800	879700
2	1	329	2000	4766	13130	11850	481900
3	2	305	1404	2975	9667	8213	565600

Analysis

```
> summary(m2)

Call:
lm(formula = log(firstPaint) ~ startup)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9963 -0.9414 -0.0794  0.8372  5.2357

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.45163    0.02438 346.700  < 2e-16 ***
startup1     0.09237    0.04193   2.203   0.0277 *
startup2    -0.29425    0.04766  -6.175 7.17e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.257 on 4954 degrees of freedom
  (17 observations deleted due to missingness)
Multiple R-squared: 0.01118,Adjusted R-squared: 0.01078
F-statistic: 28.01 on 2 and 4954 DF,  p-value: 8.007e-13
```

The p-value of 7.17e-10 means that we see a significant difference between startup values of 0 and 2 with regards to log(firstPaint). The coefficient estimate of -0.29425 means that approximately startup2 have values of firstPaint (on the original scale) which are 25% less than startup0 values. To get this simply transform the coefficient: $(1 - \exp(-0.29425) = 0.25491)$.

Note that this analysis is of an observational study. Here we are estimating an association and are not estimating a causal effect. That is to say making startup2 the default wouldn't necessary decrease firstPaint by 25%. What we can say is that in the past observations with a startup value of 2 had 25% lower firstPaint values than observations with a startup value of 0. I tried to do causal inference with this data but for now the problem seems to be intractable. This is due to the sheer number of variables (some 250 observed variables for the time period I looked at) and the large amount of missing values. That is not to say that causal inference is impossible but more thought needs to go into this and perhaps some domain knowledge in order to whittle down the large set of variables.