R Notebook For John

Contents

Load packages

```r
#install.packages("randomForest")
#install.packages("prediction")
#install.packages("class")
#install.packages("rpart")
#install.packages("olsrr")
#install.packages("broom")
#install.packages("modelr")
#install.packages("lme4")
#install.packages("lmerTest")


library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 3.5.1
```

```r
library(prediction)
```

```
## Warning: package 'prediction' was built under R version 3.5.1
```

```r
library(class)
```

```
## Warning: package 'class' was built under R version 3.5.1
```

```r
library(rpart)
```

```
## Warning: package 'rpart' was built under R version 3.5.1
```

```
library(olsrr) #John, load this package. It is way better for linear models. no need to separately get
```

```
## Warning: package 'olsrr' was built under R version 3.5.1
```

```
library(tidyverse) #Also, tidyverse is an essentail. makes manipulating and reading data a easy (takes
```

```
## Warning: package 'tidyverse' was built under R version 3.5.1

## Warning: package 'ggplot2' was built under R version 3.5.1
```

```
library(broom) #I forget what this is for! but use it anyways!
```

```
## Warning: package 'broom' was built under R version 3.5.1
```

```
library(modelr) #For hierarchical modeling
```

```
## Warning: package 'modelr' was built under R version 3.5.1
```

```
library(lme4) #For Hierarchical modeling
```

```
## Warning: package 'lme4' was built under R version 3.5.1
```

```
library(lmerTest) # <- gives hierarchical modeling p-values (There is a reason it's NOT with lme4)
```

```
## Warning: package 'lmerTest' was built under R version 3.5.1
```

Load data

```
X2017_All <- read_csv("John Dataset.csv") # <-  from tidyverse package
```

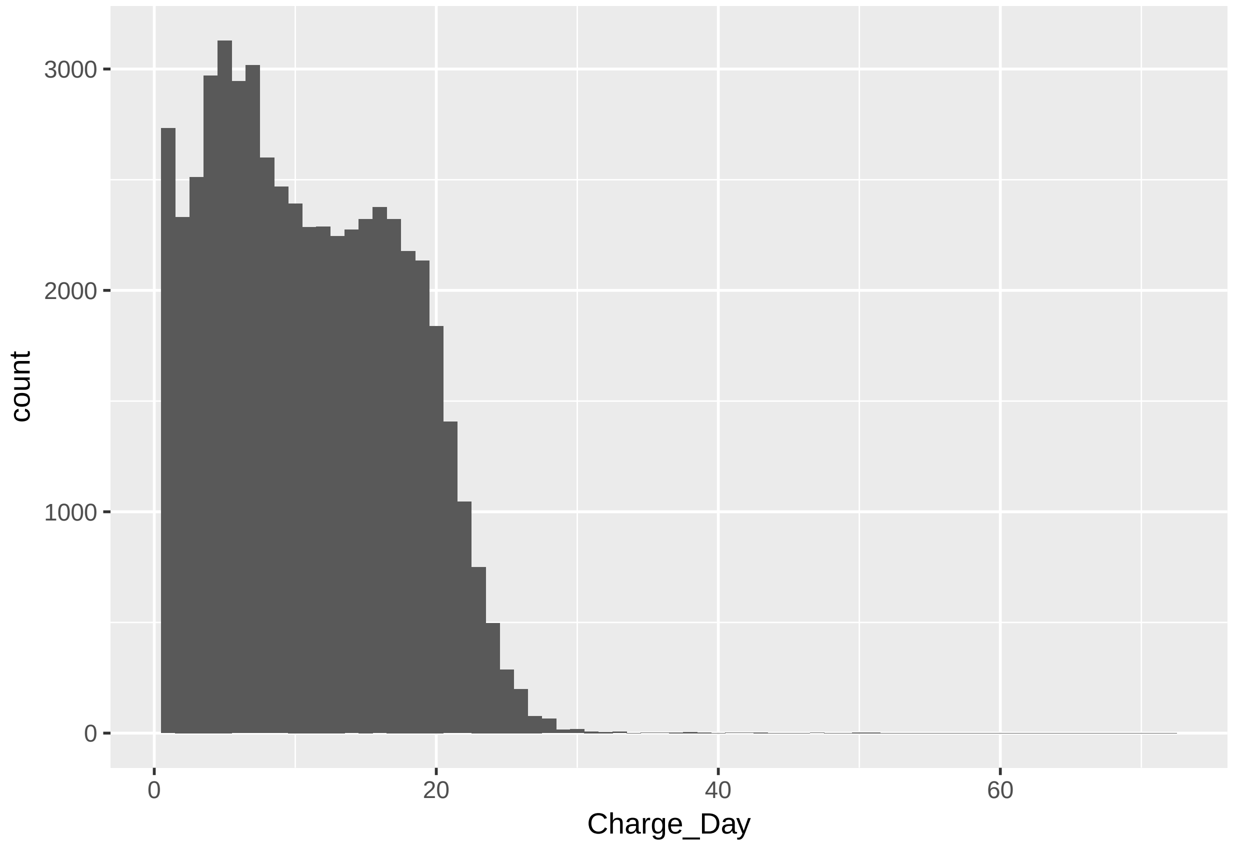
Summary stats

```
#check for outliers and set filter for interquartiles for one model
summary(X2017_All)
```

```
##  Provider_Name       Week_Number      Day_Number      Charge_Day
##  Length:53783       Min.   : 1.00   Min.   :1.000   Min.   : 1.00
##  Class :character   1st Qu.: 6.00   1st Qu.:3.000   1st Qu.: 5.00
##  Mode  :character   Median :13.00   Median :4.000   Median :10.00
##                     Mean   :18.63   Mean   :3.954   Mean   :11.01
##                     3rd Qu.:31.00   3rd Qu.:5.000   3rd Qu.:16.00
##                     Max.   :53.00   Max.   :7.000   Max.   :72.00
##    Pred Value        Pred Z               CDZ             Charge_Week
##  Min.   : 3.139   Min.   :-1.202059   Min.   :-1.528929   Min.   :  1.00
##  1st Qu.: 6.832   1st Qu.:-0.637769   1st Qu.:-0.917701   1st Qu.: 20.00
##  Median :10.524   Median :-0.073631   Median :-0.153667   Median : 38.00
##  Mean   :11.006   Mean   : 0.000001   Mean   : 0.000001   Mean   : 40.17
##  3rd Qu.:14.864   3rd Qu.: 0.589641   3rd Qu.: 0.763174   3rd Qu.: 59.00
##  Max.   :55.818   Max.   : 6.847584   Max.   : 9.320356   Max.   :133.00
##       CWZ              Year_CD          Diff
##  Min.   :-1.61915   Min.   :2017   Min.   :-16.182346
##  1st Qu.:-0.83369   1st Qu.:2017   1st Qu.: -1.336988
##  Median :-0.08956   Median :2017   Median :  0.136645
##  Mean   : 0.00000   Mean   :2017   Mean   :  0.000001
##  3rd Qu.: 0.77858   3rd Qu.:2018   3rd Qu.:  1.424091
##  Max.   : 3.83774   Max.   :2018   Max.   :  3.047649
##     All pos
##  Min.   : 0.000088
##  1st Qu.: 0.754380
##  Median : 1.410687
##  Mean   : 1.450781
##  3rd Qu.: 2.074771
##  Max.   :16.182346
```

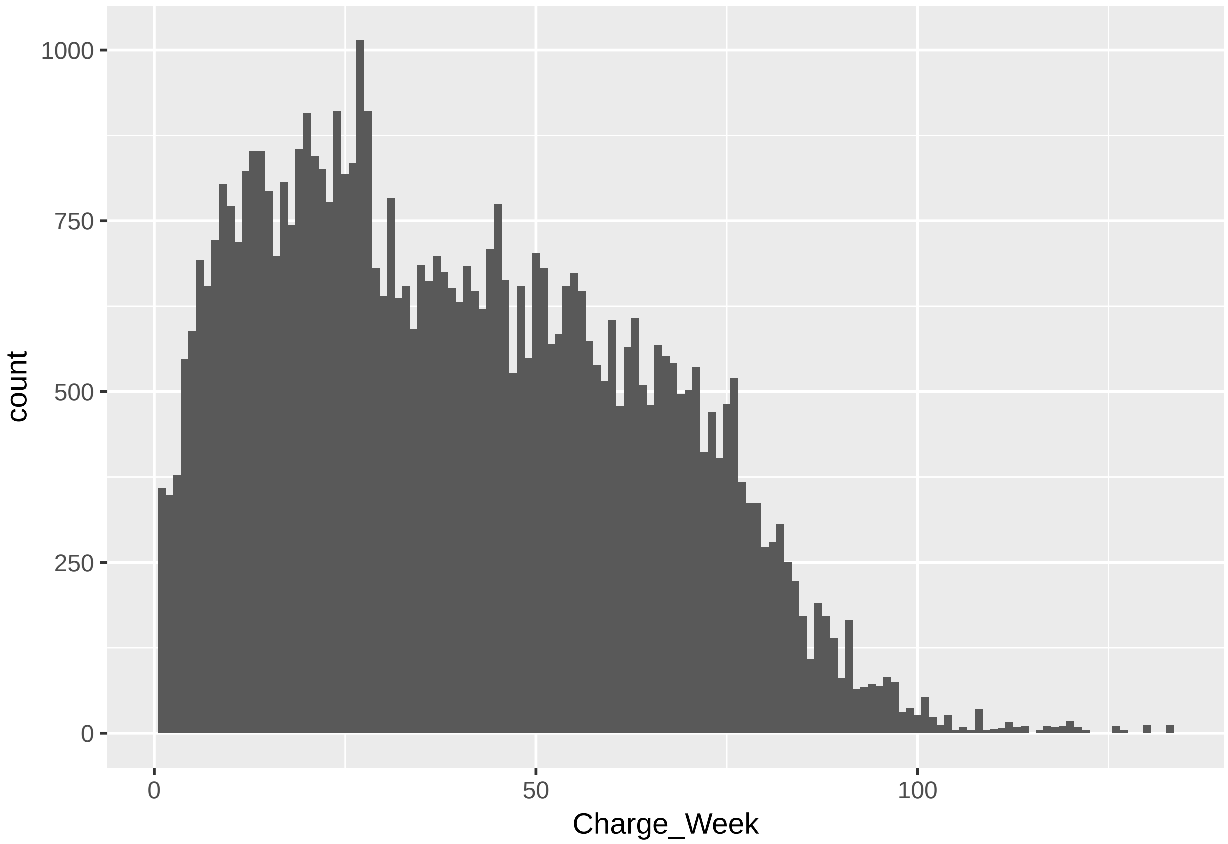
Charge_Day Histogram

```
ggplot(data = X2017_All, aes(x = Charge_Day)) +
  geom_histogram(binwidth = 1)
```



Charge Week Histogram

```
ggplot(data = X2017_All, aes(x = Charge_Week)) +
  geom_histogram(binwidth = 1)
```



Non-standardized model

We don't need to standardize this because R does this for us! Keep it regular!

```
#######################################################################
###Main Model, seems to be working best when I don't remove outliers###
#######################################################################
All<- lm(Charge_Day ~  Day_Number+ Week_Number  + Charge_Week , data = X2017_All)
ols_regress(All) # <- from olsrr
```

```
##                          Model Summary
## -------------------------------------------------------------
## R                        0.741       RMSE               4.394
## R-Squared                0.549       Coef. Var         39.924
## Adj. R-Squared           0.549       MSE               19.306
## Pred R-Squared           0.549       MAE                3.291
## -------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                      ANOVA
## ----------------------------------------------------------------------------
##                     Sum of
##                    Squares           DF     Mean Square        F          Sig.
## -----------------------------------------------------------------------------
## Regression    1265074.672            3      421691.557     21842.192    0.0000
## Residual      1038272.621        53779          19.306
## Total         2303347.293        53782
## -----------------------------------------------------------------------------
##
##                                    Parameter Estimates
## -----------------------------------------------------------------------------------------
##       model      Beta    Std. Error    Std. Beta       t        Sig      lower     upper
## -----------------------------------------------------------------------------------------
## (Intercept)     3.473         0.068                  51.303    0.000     3.340     3.605
##  Day_Number    -0.139         0.013       -0.030    -10.446    0.000    -0.165    -0.113
## Week_Number     0.001         0.001        0.003      1.136    0.256    -0.001     0.004
## Charge_Week     0.201         0.001        0.741    255.963    0.000     0.199     0.202
## -----------------------------------------------------------------------------------------
```

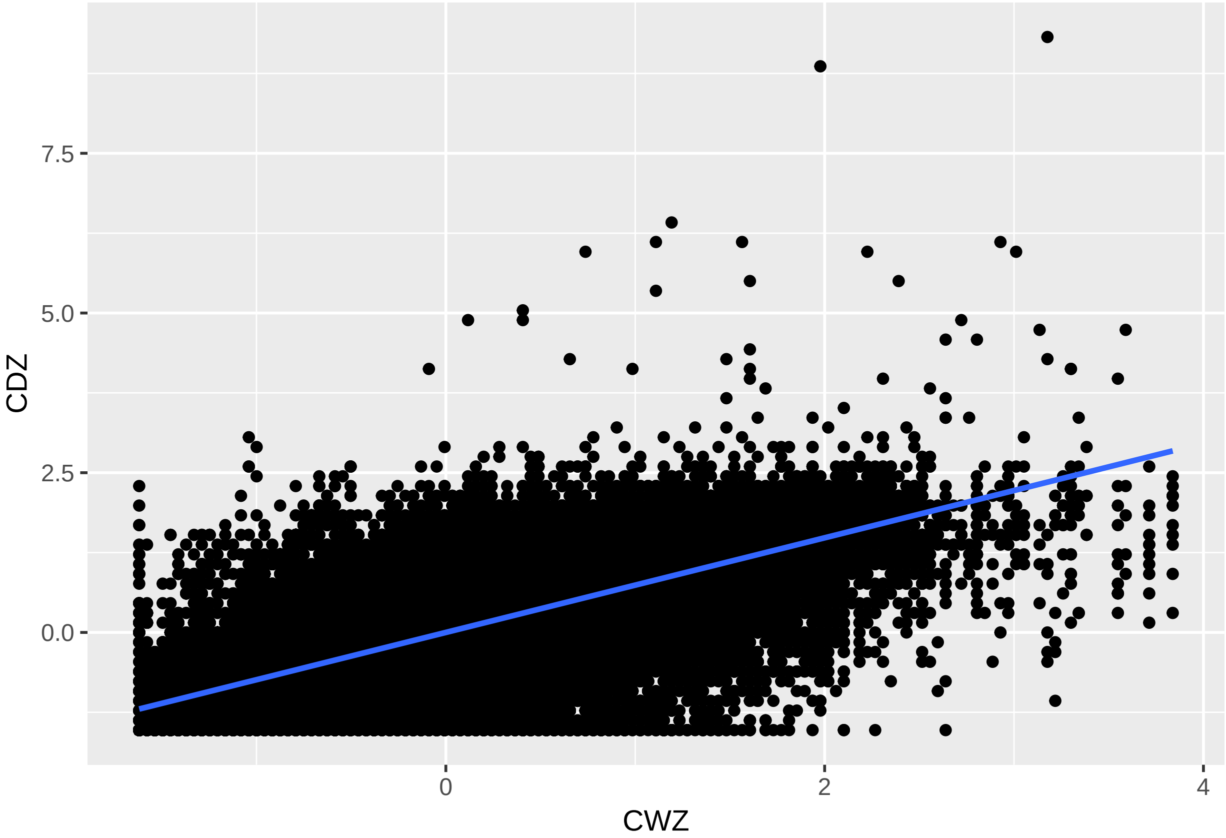
```
#summary(All)
#confint(All)
```

```
#######################################################################
###Main Model, seems to be working best when I don't remove outliers###
#######################################################################
All<- lm(CDZ ~  Day_Number+ Week_Number  + CWZ , data = X2017_All)
ols_regress(All) # <- from olsrr
```

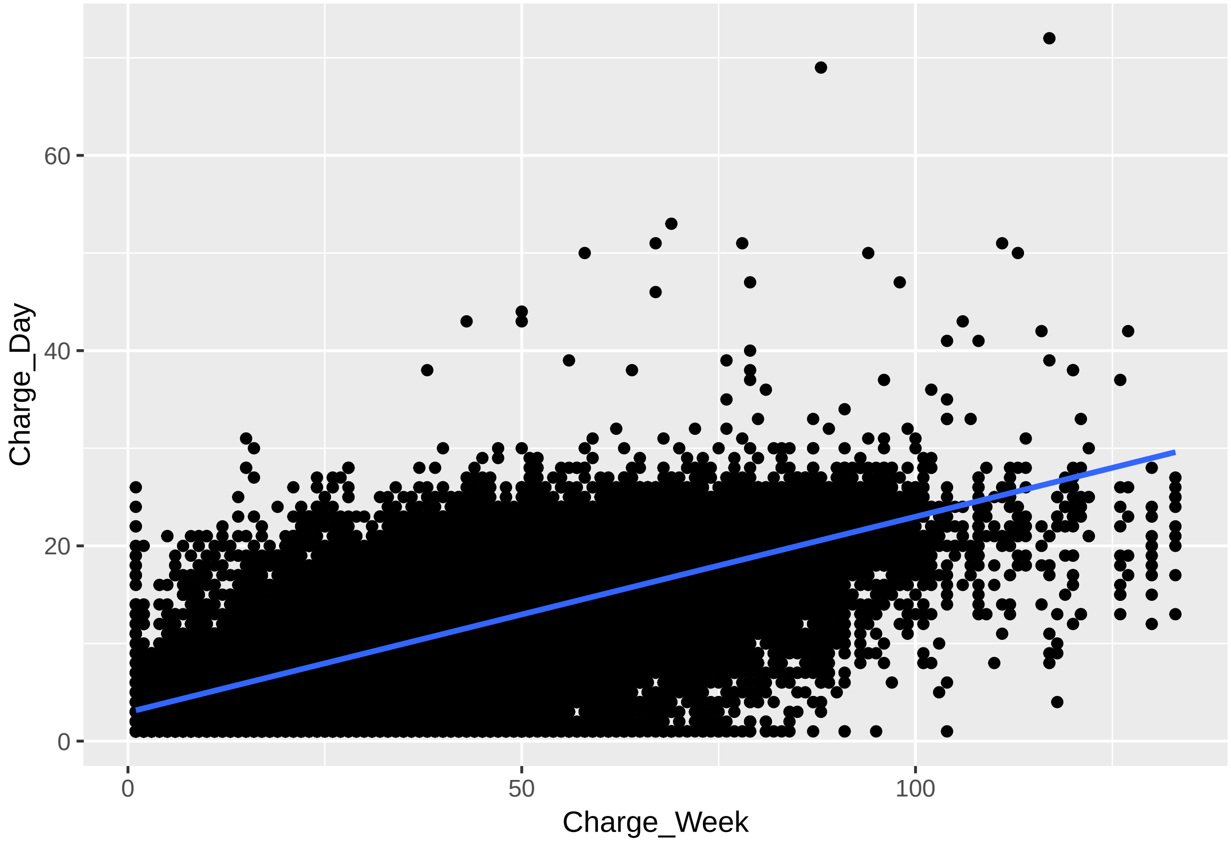
```
##                             Model Summary
## ---------------------------------------------------------------------
## R                       0.741       RMSE                       0.671
## R-Squared               0.549       Coef. Var          117151664.643
## Adj. R-Squared          0.549       MSE                        0.451
## Pred R-Squared          0.549       MAE                        0.503
## ---------------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                    ANOVA
## ---------------------------------------------------------------------------
##                  Sum of
##                 Squares           DF    Mean Square        F          Sig.
## ---------------------------------------------------------------------------
## Regression    29539.403            3       9846.468    21842.192    0.0000
## Residual      24243.591        53779          0.451
## Total         53782.993        53782
## ---------------------------------------------------------------------------
##
##                                    Parameter Estimates
## -----------------------------------------------------------------------------------------
##       model      Beta    Std. Error    Std. Beta       t        Sig      lower     upper
## -----------------------------------------------------------------------------------------
## (Intercept)     0.080         0.009                   8.617    0.000     0.062     0.098
##  Day_Number    -0.021         0.002       -0.030    -10.446    0.000    -0.025    -0.017
## Week_Number     0.000         0.000        0.003      1.136    0.256     0.000     0.001
##         CWZ     0.741         0.003        0.741    255.963    0.000     0.736     0.747
## -----------------------------------------------------------------------------------------
```

```
#summary(All)
#confint(All)
```

```
ggplot(data = X2017_All, aes(x = CWZ, y = CDZ)) +
  geom_point() +
  geom_smooth(method = "lm")
```



```
ggplot(data = X2017_All, aes(x = Charge_Week, y = Charge_Day)) +
  geom_point() +
  geom_smooth(method = "lm")
```



correlation

Proof you don't need to standardize

```
cor(X2017_All$Charge_Day, X2017_All$Charge_Week)
```

```
## [1] 0.7404776
```

```
cor(X2017_All$CDZ, X2017_All$CWZ)
```

```
## [1] 0.7404776
```

Let's try hierarchical modeling...

```
X2017_All <- X2017_All %>%
  arrange(Week_Number, Day_Number) %>%
  mutate(week.f = factor(Week_Number),
         day.f = factor(Day_Number),
         provider.f = factor(Provider_Name)) %>%
  arrange(Provider_Name, Week_Number, Day_Number)
```

Fixed effects only

```
mod1 = lmer(Charge_Day ~ 1 + Charge_Week + (1|provider.f), REML = TRUE, data = X2017_All)
summary(mod1)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: Charge_Day ~ 1 + Charge_Week + (1 | provider.f)
##    Data: X2017_All
##
## REML criterion at convergence: 297294.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.9237 -0.5051  0.0086  0.5481 14.8639
##
## Random effects:
##  Groups     Name        Variance Std.Dev.
##  provider.f (Intercept) 12.33    3.512
##  Residual               14.36    3.789
## Number of obs: 53783, groups:  provider.f, 309
##
## Fixed effects:
##              Estimate Std. Error        df t value Pr(>|t|)
## (Intercept) 6.676e+00  2.096e-01 3.065e+02   31.85   <2e-16 ***
## Charge_Week 9.766e-02  1.361e-03 5.137e+04   71.74   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr)
## Charge_Week -0.214
```

include week & day as fixed effects

This is a large dataset… it might take a while

```
mod2 = lmer(Charge_Day ~ 1 + Charge_Week + Week_Number + Day_Number +(1|provider.f), REML = TRUE, data =
summary(mod2)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: Charge_Day ~ 1 + Charge_Week + Week_Number + Day_Number + (1 |
##     provider.f)
##    Data: X2017_All
##
## REML criterion at convergence: 297233.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.8793 -0.5077  0.0083  0.5485 14.9323
##
## Random effects:
```

```
##  Groups     Name        Variance Std.Dev.
##  provider.f (Intercept) 12.36    3.516
##  Residual               14.33    3.786
## Number of obs: 53783, groups:  provider.f, 309
##
## Fixed effects:
##               Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)  7.026e+00  2.159e-01  3.431e+02  32.547   <2e-16 ***
## Charge_Week  9.775e-02  1.361e-03  5.139e+04  71.838   <2e-16 ***
## Week_Number  2.726e-03  1.109e-03  5.355e+04   2.457    0.014 *
## Day_Number  -1.026e-01  1.196e-02  5.355e+04  -8.580   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) Chrg_W Wk_Nmb
## Charge_Week -0.203
## Week_Number -0.095 -0.015
## Day_Number  -0.216 -0.012  0.004
```

Problem with this model. . . .

```
mod3 = lmer(Charge_Day ~ 1 + Charge_Week + Week_Number + Day_Number +(1+ Charge_Week |provider.f), REML
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, : Model is nearly unide
##  - Rescale variables?
```

```
summary(mod3)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: Charge_Day ~ 1 + Charge_Week + Week_Number + Day_Number + (1 +
##     Charge_Week | provider.f)
##    Data: X2017_All
##
## REML criterion at convergence: 296485.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.7332 -0.5000  0.0150  0.5485 14.0934
##
## Random effects:
##  Groups     Name        Variance  Std.Dev. Corr
##  provider.f (Intercept) 18.614390 4.31444
##             Charge_Week  0.002571 0.05071  -0.68
##  Residual               14.040546 3.74707
## Number of obs: 53783, groups:  provider.f, 309
##
## Fixed effects:
##               Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)  6.586e+00  2.617e-01  3.134e+02  25.169  < 2e-16 ***
## Charge_Week  1.245e-01  3.648e-03  1.931e+02  34.129  < 2e-16 ***
## Week_Number  3.477e-03  1.116e-03  5.320e+04   3.115  0.00184 **
## Day_Number  -1.038e-01  1.186e-02  5.338e+04  -8.756  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) Chrg_W Wk_Nmb
## Charge_Week -0.630              
## Week_Number -0.081 -0.002       
## Day_Number  -0.177 -0.007  0.005
## convergence code: 0
## Model is nearly unidentifiable: very large eigenvalue
##  - Rescale variables?
```