Linear regression in R

Erin Shellman

April 13 & 20, 2015

Contents

Linear regression

In this tutorial we'll learn:

- how to `merge` datasets
- how to fit linear regression models
- how to split data into test and train sets
- how to tune our models and select features

Data preparation

We're working with the Capital Bikeshare again this week, so start by reading in *usage*, *weather*, *stations*.

```
library(dplyr)
library(ggplot2)
library(lubridate)

usage = read.delim('usage_2012.tsv',
                   sep = '\t',
                   header = TRUE)

weather = read.delim('daily_weather.tsv',
                   sep = '\t',
                   header = TRUE)

stations = read.delim('stations.tsv',
                   sep = '\t',
                   header = TRUE)
```

Merging data

We have three related datasets to work with, but we can't really get started until they're combined. Let's start with *usage* and *weather*. The *usage* dataframe is at the resolution of the hour, while the *weather* data are at the resolution of a day, so we know we're going to have to either duplicate or compress data to merge. I vote compress, let's summarize!

```r
head(usage)
```

```
##   bike_id          time_start            time_end duration_mins
## 1  W01412 2012-01-01 00:04:00 2012-01-01 00:11:00             7
## 2  W00524 2012-01-01 00:10:00 2012-01-01 00:29:00            19
## 3  W00235 2012-01-01 00:10:00 2012-01-01 00:29:00            19
## 4  W00864 2012-01-01 00:15:00 2012-01-01 00:23:00             8
## 5  W00995 2012-01-01 00:15:00 2012-01-01 00:23:00             8
## 6  W00466 2012-01-01 00:17:00 2012-01-01 00:23:00             6
##                         station_start             station_end  cust_type
## 1         7th & R St NW / Shaw Library           7th & T St NW Registered
## 2       Georgia & New Hampshire Ave NW    16th & Harvard St NW     Casual
## 3       Georgia & New Hampshire Ave NW    16th & Harvard St NW Registered
## 4                        14th & V St NW Park Rd & Holmead Pl NW Registered
## 5                   11th & Kenyon St NW           7th & T St NW Registered
## 6 Court House Metro / 15th & N Uhle St    Lynn & 19th St North Registered
```

```r
custs_per_day =
  usage %>%
    group_by(time_start = as.Date(time_start), station_start, cust_type) %>%
    summarize(no_rentals = n(),
              duration_mins = mean(duration_mins, na.rm = TRUE))

head(custs_per_day)
```

```
## Source: local data frame [6 x 5]
## Groups: time_start, station_start
##
##   time_start                 station_start  cust_type no_rentals
## 1 2012-01-01           10th & Monroe St NE Registered         10
## 2 2012-01-01                10th & U St NW     Casual          8
## 3 2012-01-01                10th & U St NW Registered         50
## 4 2012-01-01 10th St & Constitution Ave NW     Casual         34
## 5 2012-01-01 10th St & Constitution Ave NW Registered         20
## 6 2012-01-01                11th & H St NE     Casual          4
## Variables not shown: duration_mins (dbl)
```

Perfection, now we can merge! What's the key?

```r
# make sure we have consistent date formats
custs_per_day$time_start = ymd(custs_per_day$time_start)
weather$date = ymd(weather$date)

# then merge. see ?merge for more details about the function
weather_rentals = merge(custs_per_day, weather,
```

```r
                        by.x = 'time_start', by.y = 'date')

# check dimensions after to make sure they are what you expect
dim(custs_per_day)
```

```
## [1] 99356     5
```

```r
dim(weather)
```

```
## [1] 366  15
```

```r
dim(weather_rentals)
```

```
## [1] 99356    19
```

```r
head(weather_rentals)
```

```
##   time_start                 station_start  cust_type no_rentals
## 1 2012-01-01           10th & Monroe St NE Registered         10
## 2 2012-01-01                10th & U St NW     Casual          8
## 3 2012-01-01                10th & U St NW Registered         50
## 4 2012-01-01 10th St & Constitution Ave NW     Casual         34
## 5 2012-01-01 10th St & Constitution Ave NW Registered         20
## 6 2012-01-01                11th & H St NE     Casual          4
##   duration_mins weekday season_code season_desc is_holiday is_work_day
## 1      16.40000       0           1      Spring          0           0
## 2      16.25000       0           1      Spring          0           0
## 3      10.00000       0           1      Spring          0           0
## 4      20.29412       0           1      Spring          0           0
## 5      14.20000       0           1      Spring          0           0
## 6      10.00000       0           1      Spring          0           0
##   weather_code                                     weather_desc temp
## 1            1 Clear, Few clouds, Partly cloudy, Partly cloudy 0.37
## 2            1 Clear, Few clouds, Partly cloudy, Partly cloudy 0.37
## 3            1 Clear, Few clouds, Partly cloudy, Partly cloudy 0.37
## 4            1 Clear, Few clouds, Partly cloudy, Partly cloudy 0.37
## 5            1 Clear, Few clouds, Partly cloudy, Partly cloudy 0.37
## 6            1 Clear, Few clouds, Partly cloudy, Partly cloudy 0.37
##   subjective_temp humidity windspeed no_casual_riders no_reg_riders
## 1        0.375621   0.6925  0.192167              686          1608
## 2        0.375621   0.6925  0.192167              686          1608
## 3        0.375621   0.6925  0.192167              686          1608
## 4        0.375621   0.6925  0.192167              686          1608
## 5        0.375621   0.6925  0.192167              686          1608
## 6        0.375621   0.6925  0.192167              686          1608
##   total_riders
## 1         2294
## 2         2294
## 3         2294
## 4         2294
## 5         2294
## 6         2294
```

Great, now we want to merge on the last dataset, *stations*. What is the key to link *weather_rentals* with *stations*?

```
final_data = merge(weather_rentals, stations,
                   by.x = 'station_start', by.y = 'station')
dim(final_data)
```

```
## [1] 98634   154
```

```
dim(weather_rentals)
```

```
## [1] 99356    19
```

```
head(final_data[, 1:30])
```

```
##   station_start time_start  cust_type no_rentals duration_mins weekday
## 1 10th & E St NW 2012-07-25     Casual          8      82.37500       3
## 2 10th & E St NW 2012-07-25 Registered         32      13.28125       3
## 3 10th & E St NW 2012-11-13 Subscriber         19      11.73684       2
## 4 10th & E St NW 2012-09-25 Registered         41      12.29268       2
## 5 10th & E St NW 2012-08-09 Registered         34      13.61765       4
## 6 10th & E St NW 2012-11-22 Subscriber          7      12.14286       4
##   season_code season_desc is_holiday is_work_day weather_code
## 1           3        Fall          0           1            1
## 2           3        Fall          0           1            1
## 3           4      Winter          0           1            2
## 4           4      Winter          0           1            1
## 5           3        Fall          0           1            1
## 6           4      Winter          1           0            1
##                                                      weather_desc     temp
## 1             Clear, Few clouds, Partly cloudy, Partly cloudy 0.724167
## 2             Clear, Few clouds, Partly cloudy, Partly cloudy 0.724167
## 3 Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist 0.343333
## 4             Clear, Few clouds, Partly cloudy, Partly cloudy 0.550000
## 5             Clear, Few clouds, Partly cloudy, Partly cloudy 0.755833
## 6             Clear, Few clouds, Partly cloudy, Partly cloudy 0.340000
##   subjective_temp humidity windspeed no_casual_riders no_reg_riders
## 1        0.654054 0.450000 0.1648000             1383          6790
## 2        0.654054 0.450000 0.1648000             1383          6790
## 3        0.323225 0.662917 0.3420460              327          3767
## 4        0.544179 0.570000 0.2363210              845          6693
## 5        0.699508 0.620417 0.1561000             1196          6090
## 6        0.350371 0.580417 0.0528708              955          1470
##   total_riders  id terminal_name      lat      long no_bikes
## 1         8173 199         31256 38.89591 -77.02606        6
## 2         8173 199         31256 38.89591 -77.02606        6
## 3         4094 199         31256 38.89591 -77.02606        6
## 4         7538 199         31256 38.89591 -77.02606        6
## 5         7286 199         31256 38.89591 -77.02606        6
## 6         2425 199         31256 38.89591 -77.02606        6
##   no_empty_docks fast_food parking restaurant convenience post_office
## 1              8         5       2         16           0           1
```

```
## 2                 8          5       2          16            0           1
## 3                 8          5       2          16            0           1
## 4                 8          5       2          16            0           1
## 5                 8          5       2          16            0           1
## 6                 8          5       2          16            0           1
```

```
# probably want to save this now!
write.table(final_data,
            'bikeshare_modeling_data.tsv',
            row.names = FALSE, sep = '\t')

# rename to something more convenient and remove from memory
data = final_data
rm(final_data)
```

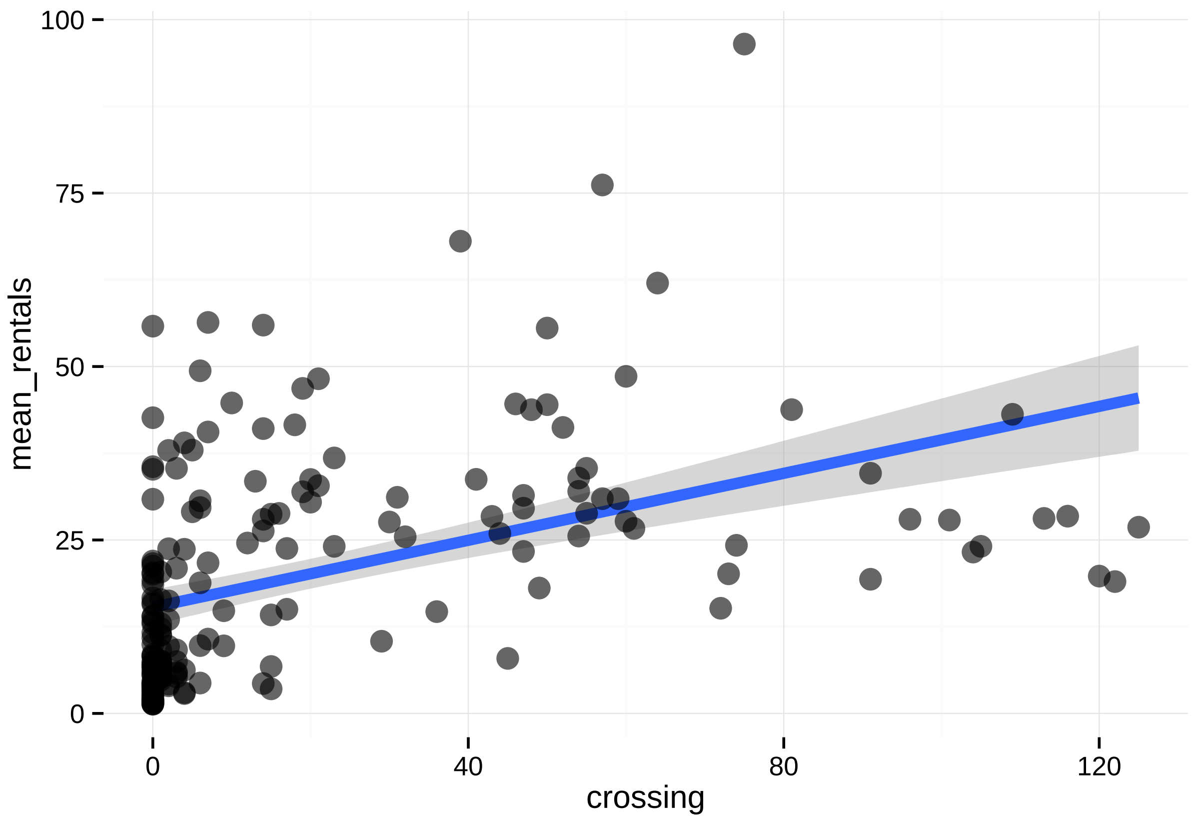
The `lm()` function

The function for creating a linear model in R is `lm()` and the primary arguments are *formula* and *data*. Formulas in R are a little funny, instead of an = sign, they are expressed with a ~. Let's fit the model we saw in the lecture notes: $rentals = \beta_0 + \beta_1 * crossing$. There's a little snag we have to take care of first. Right now we've got repeated measures *i.e.* one measurement per day, so we need to aggregate again this time over date.

```
rentals_crossing =
  data %>%
    group_by(station_start) %>%
    summarize(mean_rentals = mean(no_rentals),
              crossing = mean(crossing))

head(rentals_crossing)
```

```
## Source: local data frame [6 x 3]
##
##                   station_start mean_rentals crossing
## 1                10th & E St NW    19.003003      122
## 2           10th & Monroe St NE     7.580517        1
## 3                10th & U St NW    37.954876        5
## 4 10th St & Constitution Ave NW    28.430362      116
## 5                11th & H St NE    20.121875       73
## 6           11th & Kenyon St NW    33.718331       20
```

```
# plot it
ggplot(rentals_crossing, aes(x = crossing, y = mean_rentals)) +
  geom_smooth(method = 'lm', size = 2) +
  geom_point(size = 4, alpha = 0.60) +
  theme_minimal()
```




```r
model = lm(mean_rentals ~ crossing, data = rentals_crossing)

# view what is returned in the lm object
attributes(model)
```

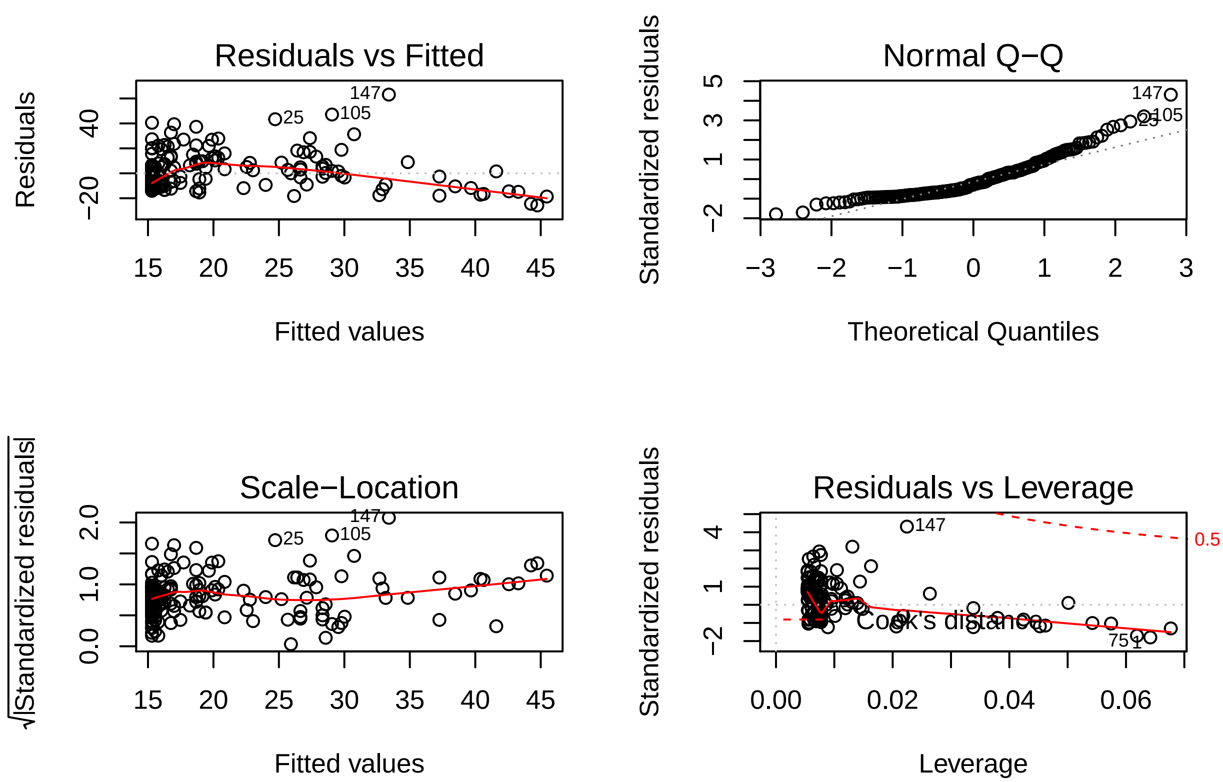
```
## $names
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
##
## $class
## [1] "lm"
```

```r
# get model output
summary(model)
```

```
##
## Call:
## lm(formula = mean_rentals ~ crossing, data = rentals_crossing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.735 -10.767  -4.190   6.755  63.079
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.30402    1.29989  11.773  < 2e-16 ***
```

```
## crossing     0.24127    0.03524   6.846 1.11e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.8 on 183 degrees of freedom
## Multiple R-squared:  0.2039, Adjusted R-squared:  0.1996
## F-statistic: 46.87 on 1 and 183 DF,  p-value: 1.109e-10
```

```
# print model diagnostics
par(mfrow = c(2, 2))
plot(model)
```



The `attributes()` function can be called on just about any object in R and it returns a list of all the things inside. It's a great way to explore objects and see what values are contained inside that could be used in other analysis. For example, extracting the residuals via `model$residuals` is useful if we want to print diagnostic plots like those above.

When we run `summary()` on the `lm` object, we see the results. The *Call* section just prints back the model specification, and the *Residuals* section contains a summary of the distribution of the errors. The fun stuff is in the *Coefficients* section. In the first row contains the covariate names followed by their estimates, standard errors, t- and p-values. Our model ends up being `rentals = 15 + 0.24(crosswalks)` which means that the average number of rentals when there are no crosswalks is 15, and the average increases by 1 rental for every four additional crosswalks.

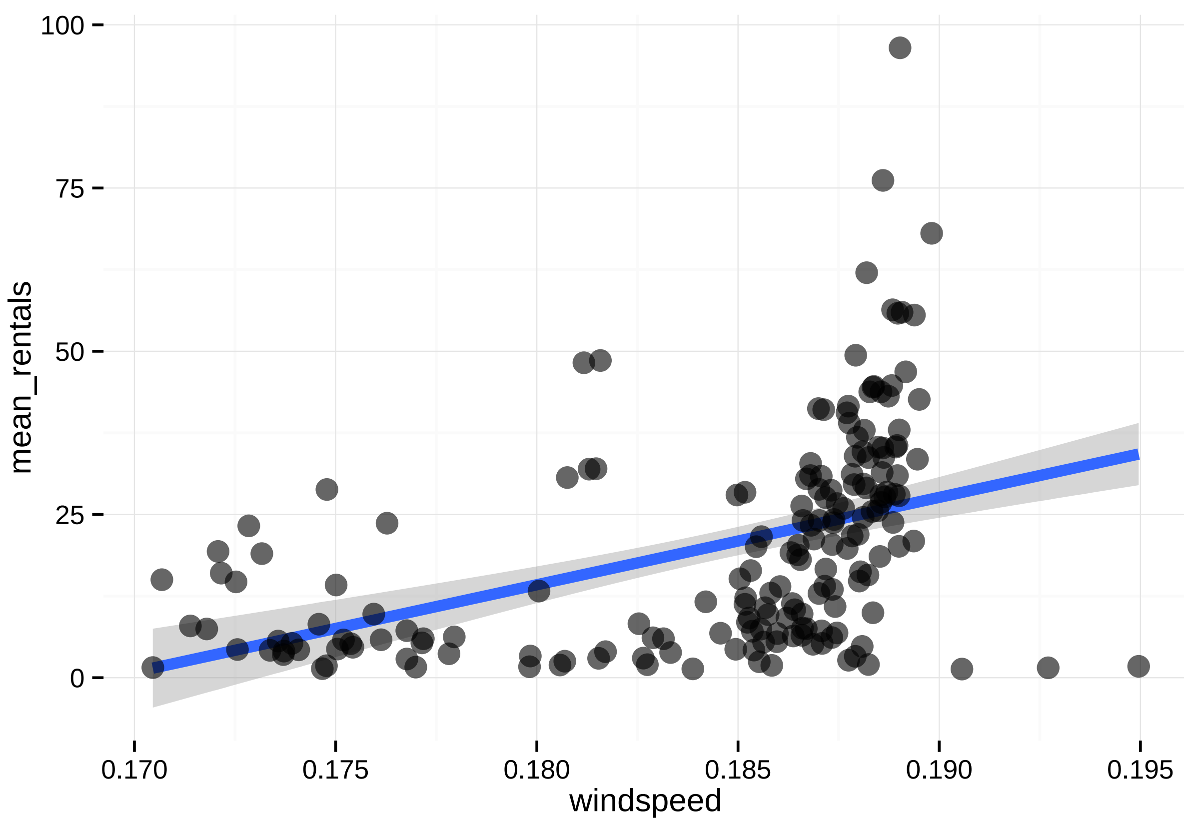
We can fit regressions with multiple covariates the same way:

```
# lets include windspeed this time
rentals_multi =
  data %>%
    group_by(station_start) %>%
```

```
    summarize(mean_rentals = mean(no_rentals),
              crossing = mean(crossing),
              windspeed = mean(windspeed))

head(rentals_multi)
```

```
## Source: local data frame [6 x 4]
##
##                     station_start mean_rentals crossing windspeed
## 1                 10th & E St NW    19.003003      122 0.1731664
## 2            10th & Monroe St NE     7.580517        1 0.1866016
## 3                 10th & U St NW    37.954876        5 0.1890061
## 4 10th St & Constitution Ave NW    28.430362      116 0.1886993
## 5                 11th & H St NE    20.121875       73 0.1889982
## 6            11th & Kenyon St NW    33.718331       20 0.1882405
```

```
ggplot(rentals_multi, aes(x = windspeed, y = mean_rentals)) +
  geom_smooth(method = 'lm', size = 2) +
  geom_point(size = 4, alpha = 0.60) +
  theme_minimal()
```




```
model = lm(mean_rentals ~ crossing + windspeed, data = rentals_multi)
summary(model)
```

```
##
## Call:
```

```
## lm(formula = mean_rentals ~ crossing + windspeed, data = rentals_multi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.454  -9.202  -1.752   5.080  59.203
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -200.35799   34.20198  -5.858 2.15e-08 ***
## crossing       0.21373    0.03231   6.616 3.99e-10 ***
## windspeed   1172.33663  185.81081   6.309 2.07e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.45 on 182 degrees of freedom
## Multiple R-squared:  0.3468, Adjusted R-squared:  0.3396
## F-statistic: 48.31 on 2 and 182 DF,  p-value: < 2.2e-16
```

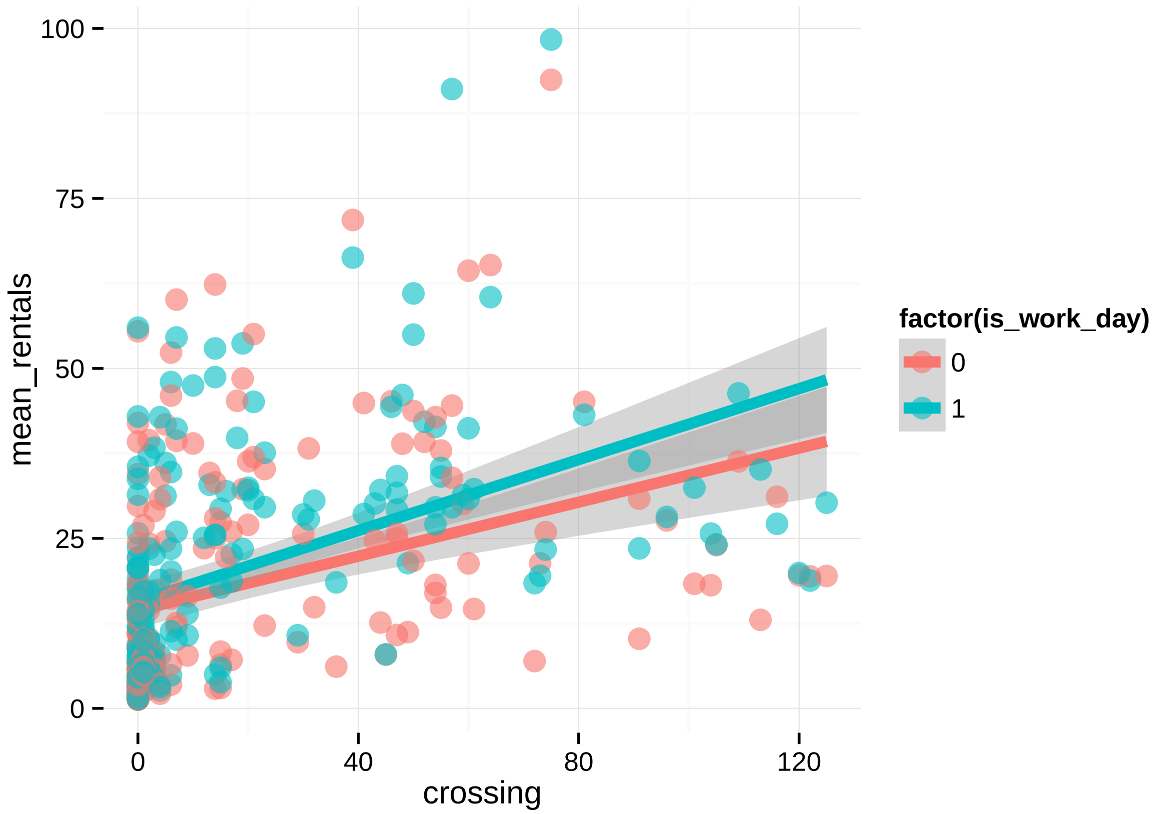
The model coefficients changed quite a lot when we added in wind speed. The intercept is now negative, and the wind speed coefficient is huge! When interpreting coefficients, it's important to keep the scale in mind. Wind speed ranges from 0.05 to 0.44 so when you multiply 1172 by 0.05 for example, you end up with about 60, which is within the range we'd expect.

Let's try one more, this time we'll include a factor variable.

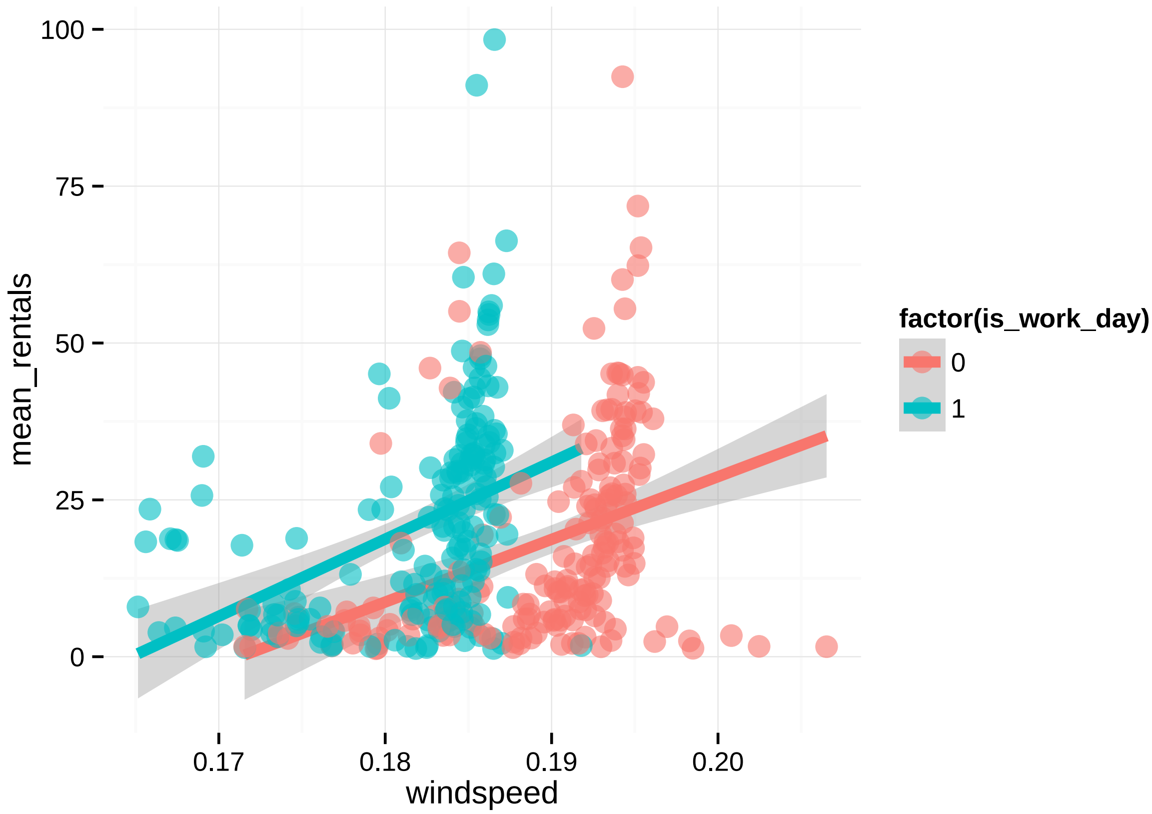
```r
rentals_multi =
  data %>%
    group_by(station_start, is_work_day) %>%
    summarize(mean_rentals = mean(no_rentals),
              crossing = mean(crossing),
              windspeed = mean(windspeed))

head(rentals_multi)
```

```
## Source: local data frame [6 x 5]
## Groups: station_start
##
##         station_start is_work_day mean_rentals crossing windspeed
## 1      10th & E St NW           0    19.416667      122 0.1858375
## 2      10th & E St NW           1    18.804444      122 0.1670843
## 3 10th & Monroe St NE           0     5.854054        1 0.1912622
## 4 10th & Monroe St NE           1     8.584906        1 0.1838902
## 5      10th & U St NW           0    41.761062        5 0.1939839
## 6      10th & U St NW           1    36.088937        5 0.1865657
```

```r
# plot crossings, colored by is_work_day
ggplot(rentals_multi,
       aes(x = crossing, y = mean_rentals, color = factor(is_work_day))) +
  geom_smooth(method = 'lm', size = 2) +
  geom_point(size = 4, alpha = 0.60) +
  theme_minimal()
```




```r
# plot windspeed, colored by is_work_day
ggplot(rentals_multi,
       aes(x = windspeed, y = mean_rentals, color = factor(is_work_day))) +
  geom_smooth(method = 'lm', size = 2) +
  geom_point(size = 4, alpha = 0.60) +
  theme_minimal()
```




```r
model = lm(mean_rentals ~ crossing + windspeed + factor(is_work_day),
           data = rentals_multi)
summary(model)
```

```
##
## Call:
## lm(formula = mean_rentals ~ crossing + windspeed + factor(is_work_day),
##     data = rentals_multi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.943  -9.728  -2.500   5.734  61.718
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -165.77396   24.38634  -6.798 4.33e-11 ***
## crossing                 0.20358    0.02448   8.316 1.81e-15 ***
## windspeed              949.26045  128.75542   7.373 1.13e-12 ***
## factor(is_work_day)1    10.05016    1.81045   5.551 5.46e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.39 on 366 degrees of freedom
## Multiple R-squared:  0.2868, Adjusted R-squared:  0.281
## F-statistic: 49.06 on 3 and 366 DF,  p-value: < 2.2e-16
```

The output looks a little funny now. There's a term called `factor(is_work_day)1`, what does that mean? Factors are category variables and their interpretation is relative to a baseline. Our factor `is_work_day` only has two levels, 0 and 1, and R sets 0 to the baseline by default. So the interpretation of that term is that we can expect about 10 additional rentals when it is a work day (*i.e.* `is_work_day == 0`) and the other variables are fixed.

The *caret* package

We'll be using the *caret* package (short for *c*lassification *a*nd *re*gression *t*raining) for model development because it integrates many modeling packages in R into one unified syntax. That means more reusable code for us! *caret* contains helper functions that provide a unified framework for data cleaning/splitting, model training, and comparison. I highly recommend the optional reading this week which provides a great overview of the *caret* package.

```
install.packages('caret', dependencies = TRUE)
library(caret)

set.seed(1234) # set a seed
```

Setting a seed in R insures that you get identical results each time you run your code. Since resampling methods are inherently probabilistic, every time we rerun them we'll get slightly different answers. Setting the seed to the same number insures that we get identical randomness each time the code is run, and that's helpful for debugging.

Train and test data

Before any analysis in this class we'll need to divide our data into train and test sets. Check out this nice overview for more details. The *training* set is typically about 75% of the data and is used for all the model development. Once we have a model we're satified with, we use our *testing* set, the other 25% to generate model predictions. Splitting the data into the two groups, train and test, generates two types of errors, in-sample and out-of-sample errors. *In-sample* errors are the errors derived from same data the model was built with. *Out-of-sample* errors are derived from measuring the error on a fresh data set. We are interested in the out-of-sample error because this quantity represents how'd we'd expect the model to perform in the future on brand new data.

Here's how to split the data with *caret*:

```
# select the training observations
in_train = createDataPartition(y = rentals_multi$mean_rentals,
                               p = 0.75, # 75% in train, 25% in test
                               list = FALSE)
head(in_train) # row indices of observations in the training set
```

```
##      Resample1
## [1,]        13
## [2,]        17
## [3,]        41
## [4,]        43
## [5,]        44
## [6,]        87
```

```
train = rentals_multi[in_train, ]
test = rentals_multi[-in_train, ]

dim(train)
```

```
## [1] 278   5
```

```
dim(test)
```

```
## [1] 92  5
```

Note: I recommend doing all data processing and aggregation steps *before* splitting out your train/test sets.

Training

Our workhorse function in the *caret* package in the `train` function. This function can be used to evaluate performance parameters, choose optimal models based on the values of those parameters, and estimate model performance. For regression we can use it in place of the `lm()` function. Here's our last regression model using the train function.

```
model_fit = train(mean_rentals ~ crossing + windspeed + factor(is_work_day),
                  data = train,
                  method = 'lm',
                  metric = 'RMSE')
print(model_fit)
```

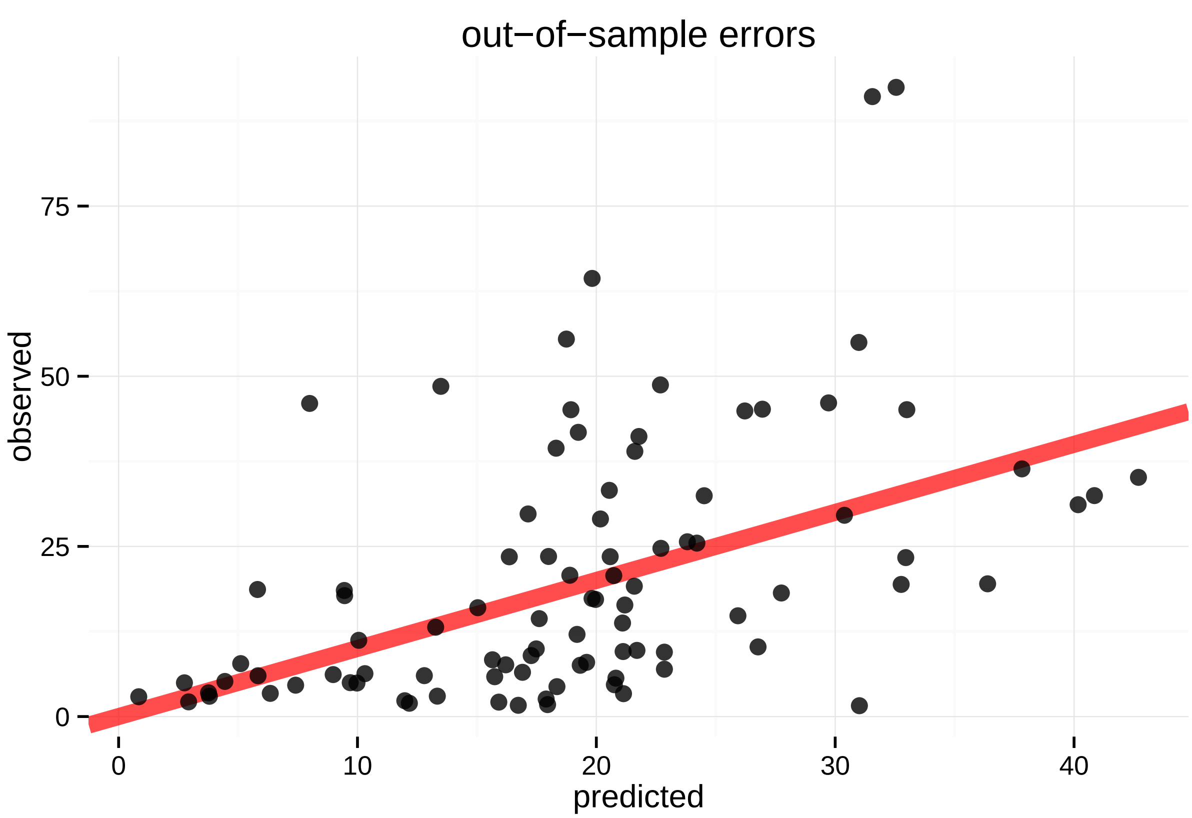
```
## Linear Regression
##
## 278 samples
##   4 predictor
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
##
## Summary of sample sizes: 278, 278, 278, 278, 278, 278, ...
##
## Resampling results
##
##   RMSE      Rsquared   RMSE SD    Rsquared SD
##   14.14664  0.2903549  0.9607718  0.05188958
##
##
```

```
summary(model_fit)
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.947  -9.053  -2.401   6.079  62.374
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -178.08113   27.19168  -6.549 2.85e-10 ***
```

```
## crossing                   0.18602    0.02736   6.800 6.54e-11 ***
## windspeed               1012.43658  143.44299   7.058 1.39e-11 ***
## `factor(is_work_day)1`    11.23121    2.02050   5.559 6.45e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.67 on 274 degrees of freedom
## Multiple R-squared:  0.3042, Adjusted R-squared:  0.2966
## F-statistic: 39.93 on 3 and 274 DF,  p-value: < 2.2e-16
```

```
# get predictions
out_of_sample_predictions = predict(model_fit, newdata = test)

# compare predictions against the observed values
errors = data.frame(predicted = out_of_sample_predictions,
                    observed = test$mean_rentals,
                    error = out_of_sample_predictions - test$mean_rentals)

# eh, not so good
ggplot(data = errors, aes(x = predicted, y = observed)) +
  geom_abline(aes(intercept = 0, slope = 1),
              size = 3, alpha = 0.70, color = 'red') +
  geom_point(size = 3, alpha = 0.80) +
  ggtitle('out-of-sample errors') +
  theme_minimal()
```



Our prediction accuracy is not so great for this model. The RMSE is about 15 which means that on average the predictions are off by about 15 rentals.

Parameter tuning

Feature Selection

Which model is the best?

Typically adding more predictors to a model will increase the R^2, so using that criteria alone will cause you to favor larger models.

Project tips

We saw how to merge the datasets together into one, but it often makes sense to do some aggregation before merging. For example, since we know *usage* needs to be aggregated and summarized to remove the date variable, it makes sense to merge *usage* with the weather data and summarized before merging on the station data. For example:

```
# we made this data frame in the merging section above
weather_rentals = merge(custs_per_day, weather,
                        by.x = 'time_start', by.y = 'date')

# group_by all the factors and summarize the continuous variables to generate
# a final data frame that can be merged by station.
model_data =
  weather_rentals %>%
    group_by(
      station_start,
      cust_type,
      weekday,
      season_code,
      is_holiday,
      is_work_day,
      weather_code) %>%
    summarize(
      rentals = mean(no_rentals),
      duration = mean(duration_mins),
      temp = mean(temp),
      subjective_temp = mean(subjective_temp),
      humidity = mean(humidity),
      windspeed = mean(windspeed))

head(model_data)
```

```
## Source: local data frame [6 x 13]
## Groups: station_start, cust_type, weekday, season_code, is_holiday, is_work_day
##
##    station_start cust_type weekday season_code is_holiday is_work_day
## 1 10th & E St NW    Casual       0           1          0           0
## 2 10th & E St NW    Casual       0           3          0           0
## 3 10th & E St NW    Casual       0           3          0           0
## 4 10th & E St NW    Casual       0           4          0           0
## 5 10th & E St NW    Casual       0           4          0           0
## 6 10th & E St NW    Casual       1           1          0           1
```

```
## Variables not shown: weather_code (int), rentals (dbl), duration (dbl),
##   temp (dbl), subjective_temp (dbl), humidity (dbl), windspeed (dbl)
```

```r
# now merge on stations
final_data = merge(model_data, stations,
  by.x = 'station_start',
  by.y = 'station')

data = final_data
rm(final_data)

# remove variables from the data that won't be used for modeling, e.g. lat/long
data_to_model =
  data %>%
    select(-station_start, -id, -terminal_name, -lat, -long)

dim(data_to_model)
```

```
## [1] 23390   143
```

```r
head(data_to_model)
```

```
##   cust_type weekday season_code is_holiday is_work_day weather_code
## 1    Casual       2           3          0           1            2
## 2    Casual       2           4          0           1            1
## 3    Casual       2           4          0           1            2
## 4    Casual       3           3          0           1            1
## 5    Casual       3           3          0           1            2
## 6    Casual       3           4          0           1            1
##   rentals duration      temp subjective_temp  humidity windspeed no_bikes
## 1   10.50 19.77941 0.6795830       0.6313440 0.7881250 0.2372475        6
## 2   10.60 28.57302 0.4898332       0.4828182 0.6340000 0.1817242        6
## 3    5.40 22.91333 0.3618334       0.3523850 0.6985836 0.2297340        6
## 4   11.00 56.17625 0.6869446       0.6442257 0.6009258 0.1478740        6
## 5   23.00 54.26087 0.7500000       0.7077170 0.6729170 0.1107000        6
## 6    7.75 28.51447 0.4507291       0.4406516 0.6083854 0.1771694        6
##   no_empty_docks fast_food parking restaurant convenience post_office
## 1              8         5       2         16           0           1
## 2              8         5       2         16           0           1
## 3              8         5       2         16           0           1
## 4              8         5       2         16           0           1
## 5              8         5       2         16           0           1
## 6              8         5       2         16           0           1
##   bicycle_parking drinking_water recycling waste_basket waste_disposal
## 1               4              0         0            0              0
## 2               4              0         0            0              0
## 3               4              0         0            0              0
## 4               4              0         0            0              0
## 5               4              0         0            0              0
## 6               4              0         0            0              0
##   cafe currency_exchange fountain ice_cream optician pharmacy
## 1    6                 0        0         0        0        0
## 2    6                 0        0         0        0        0
```

```
## 3    6                 0        0         0        0        0
## 4    6                 0        0         0        0        0
## 5    6                 0        0         0        0        0
## 6    6                 0        0         0        0        0
##   tanning_salon car_sharing alcohol bank bar club embassy food_court
## 1             0           0       0    4   1    0       0          0
## 2             0           0       0    4   1    0       0          0
## 3             0           0       0    4   1    0       0          0
## 4             0           0       0    4   1    0       0          0
## 5             0           0       0    4   1    0       0          0
## 6             0           0       0    4   1    0       0          0
##   government internal_kindergarten kindergarten place_of_worship post_box
## 1          0                     0            0                2        1
## 2          0                     0            0                2        1
## 3          0                     0            0                2        1
## 4          0                     0            0                2        1
## 5          0                     0            0                2        1
## 6          0                     0            0                2        1
##   pub vending_machine fuel grave_yard public_building school fire_station
## 1   1               0    0          0               0      0            0
## 2   1               0    0          0               0      0            0
## 3   1               0    0          0               0      0            0
## 4   1               0    0          0               0      0            0
## 5   1               0    0          0               0      0            0
## 6   1               0    0          0               0      0            0
##   nightclub atm hospital doctors theatre university clock parking_entrance
## 1         0   1        0       0       0          0     0                0
## 2         0   1        0       0       0          0     0                0
## 3         0   1        0       0       0          0     0                0
## 4         0   1        0       0       0          0     0                0
## 5         0   1        0       0       0          0     0                0
## 6         0   1        0       0       0          0     0                0
##   police cultural_center stripclub marketplace dry_cleaner
## 1      0               0         0           0           0
## 2      0               0         0           0           0
## 3      0               0         0           0           0
## 4      0               0         0           0           0
## 5      0               0         0           0           0
## 6      0               0         0           0           0
##   bicycle_repair_station office arts_centre library studio strip_club
## 1                      0      0           0       0      0          0
## 2                      0      0           0       0      0          0
## 3                      0      0           0       0      0          0
## 4                      0      0           0       0      0          0
## 5                      0      0           0       0      0          0
## 6                      0      0           0       0      0          0
##   tourist veterinary community_centre compressed_air tutor clinic dentist
## 1       1          0                0              0     0      0       0
## 2       1          0                0              0     0      0       0
## 3       1          0                0              0     0      0       0
## 4       1          0                0              0     0      0       0
## 5       1          0                0              0     0      0       0
## 6       1          0                0              0     0      0       0
##   bench cinema college parking_exit bar.restaurant car_rental coworking
```

```
## 1     0      1       0             0             0          0         0
## 2     0      1       0             0             0          0         0
## 3     0      1       0             0             0          0         0
## 4     0      1       0             0             0          0         0
## 5     0      1       0             0             0          0         0
## 6     0      1       0             0             0          0         0
##   shelter bureau_de_change food_cart school..historic. border_control
## 1       0                0         0                 0              0
## 2       0                0         0                 0              0
## 3       0                0         0                 0              0
## 4       0                0         0                 0              0
## 5       0                0         0                 0              0
## 6       0                0         0                 0              0
##   check_cashing nail_salon storage tax catering dojo tax_service
## 1             0          0       0   0        0    0           0
## 2             0          0       0   0        0    0           0
## 3             0          0       0   0        0    0           0
## 4             0          0       0   0        0    0           0
## 5             0          0       0   0        0    0           0
## 6             0          0       0   0        0    0           0
##   bus_station hospital..historic. toilets marker social_facility telephone
## 1           0                   0       0      0               0         0
## 2           0                   0       0      0               0         0
## 3           0                   0       0      0               0         0
## 4           0                   0       0      0               0         0
## 5           0                   0       0      0               0         0
## 6           0                   0       0      0               0         0
##   taxi building gym emergency_phone courthouse fitness_center townhall
## 1    0        0   0               0          0              0        0
## 2    0        0   0               0          0              0        0
## 3    0        0   0               0          0              0        0
## 4    0        0   0               0          0              0        0
## 5    0        0   0               0          0              0        0
## 6    0        0   0               0          0              0        0
##   car_wash ev_charging recycling.waste_basket sign charging_station
## 1        0           0                      0    0                0
## 2        0           0                      0    0                0
## 3        0           0                      0    0                0
## 4        0           0                      0    0                0
## 5        0           0                      0    0                0
## 6        0           0                      0    0                0
##   photography picnic_table nursing_home traffic_signals crossing
## 1           0            0            0              77      122
## 2           0            0            0              77      122
## 3           0            0            0              77      122
## 4           0            0            0              77      122
## 5           0            0            0              77      122
## 6           0            0            0              77      122
##   motorway_junction bus_stop speed_camera service stop turning_circle
## 1                 0        2            0       0    1              0
## 2                 0        2            0       0    1              0
## 3                 0        2            0       0    1              0
## 4                 0        2            0       0    1              0
## 5                 0        2            0       0    1              0
```

```
## 6                 0         2             0       0    1             0
##   elevator traffic_signals.bus_stop mini_roundabout footway street_lamp
## 1        0                        0               0       0           0
## 2        0                        0               0       0           0
## 3        0                        0               0       0           0
## 4        0                        0               0       0           0
## 5        0                        0               0       0           0
## 6        0                        0               0       0           0
##   turning_loop hotel artwork information museum sculpture hostel
## 1            0     3       0           1      1         0      0
## 2            0     3       0           1      1         0      0
## 3            0     3       0           1      1         0      0
## 4            0     3       0           1      1         0      0
## 5            0     3       0           1      1         0      0
## 6            0     3       0           1      1         0      0
##   picnic_site tour_guide attraction landmark motel guest_house gallery
## 1           0          1          0        0     0           0       0
## 2           0          1          0        0     0           0       0
## 3           0          1          0        0     0           0       0
## 4           0          1          0        0     0           0       0
## 5           0          1          0        0     0           0       0
## 6           0          1          0        0     0           0       0
```

```r
model = lm(rentals ~ ., data = data_to_model)
summary(model)
```

```
##
## Call:
## lm(formula = rentals ~ ., data = data_to_model)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.933 -11.365  -0.838   8.980 222.391
##
## Coefficients: (23 not defined because of singularities)
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -1.737e+01  1.618e+00 -10.740  < 2e-16 ***
## cust_typeRegistered       2.755e+01  2.854e-01  96.523  < 2e-16 ***
## cust_typeSubscriber       1.603e+01  4.039e-01  39.699  < 2e-16 ***
## weekday                   6.292e-01  6.487e-02   9.700  < 2e-16 ***
## season_code              -5.139e-01  1.404e-01  -3.662 0.000251 ***
## is_holiday               -3.132e+00  4.912e-01  -6.376 1.85e-10 ***
## is_work_day               2.061e+00  2.851e-01   7.231 4.94e-13 ***
## weather_code             -3.945e+00  4.162e-01  -9.479  < 2e-16 ***
## duration                  8.699e-04  4.891e-04   1.779 0.075313 .
## temp                     -1.483e+02  1.087e+01 -13.637  < 2e-16 ***
## subjective_temp           1.700e+02  1.262e+01  13.464  < 2e-16 ***
## humidity                 -7.251e-01  2.299e+00  -0.315 0.752438
## windspeed                -1.540e+01  2.692e+00  -5.719 1.08e-08 ***
## no_bikes                  8.066e-01  3.916e-02  20.597  < 2e-16 ***
## no_empty_docks            6.027e-01  3.984e-02  15.129  < 2e-16 ***
## fast_food                 4.224e-01  1.623e-01   2.602 0.009271 **
## parking                  -1.321e+00  3.253e-01  -4.062 4.89e-05 ***
## restaurant                1.273e-01  6.479e-02   1.965 0.049453 *
```

```
## convenience              -2.393e+01  2.985e+00  -8.016 1.14e-15 ***
## post_office               7.823e-02  6.808e-01   0.115 0.908520    
## bicycle_parking          -2.040e+00  5.089e-01  -4.008 6.14e-05 ***
## drinking_water            1.715e+00  3.819e-01   4.490 7.15e-06 ***
## recycling                -6.329e+00  1.719e+00  -3.681 0.000233 ***
## waste_basket              1.243e+00  2.504e-01   4.963 6.98e-07 ***
## waste_disposal                   NA         NA      NA       NA    
## cafe                      6.801e-01  2.560e-01   2.656 0.007902 ** 
## currency_exchange         3.960e+01  2.482e+00  15.957  < 2e-16 ***
## fountain                  3.233e+00  6.346e-01   5.095 3.52e-07 ***
## ice_cream                        NA         NA      NA       NA    
## optician                         NA         NA      NA       NA    
## pharmacy                  1.131e+00  6.380e-01   1.772 0.076344 .  
## tanning_salon                    NA         NA      NA       NA    
## car_sharing               2.301e+00  1.083e+00   2.124 0.033657 *  
## alcohol                   3.924e+00  2.349e+00   1.670 0.094870 .  
## bank                     -2.075e+00  1.681e-01 -12.344  < 2e-16 ***
## bar                       1.650e+00  1.466e-01  11.252  < 2e-16 ***
## club                     -1.294e+01  3.079e+00  -4.203 2.64e-05 ***
## embassy                   8.422e+00  7.531e-01  11.182  < 2e-16 ***
## food_court               -1.314e+01  1.974e+00  -6.656 2.87e-11 ***
## government                6.923e+00  3.857e+00   1.795 0.072650 .  
## internal_kindergarten     2.657e+01  4.409e+00   6.026 1.70e-09 ***
## kindergarten             -1.319e-02  4.306e-01  -0.031 0.975564    
## place_of_worship          1.489e+00  1.405e-01  10.594  < 2e-16 ***
## post_box                 -1.685e+00  5.213e-01  -3.232 0.001231 ** 
## pub                       4.257e+00  2.776e-01  15.338  < 2e-16 ***
## vending_machine                  NA         NA      NA       NA    
## fuel                      1.432e+00  8.514e-01   1.682 0.092657 .  
## grave_yard                2.614e+00  7.687e-01   3.401 0.000672 ***
## public_building           2.954e+00  6.115e-01   4.830 1.37e-06 ***
## school                    2.172e+00  1.732e-01  12.542  < 2e-16 ***
## fire_station             -2.859e-01  8.600e-01  -0.332 0.739528    
## nightclub                -1.673e+01  1.266e+00 -13.215  < 2e-16 ***
## atm                      -4.002e+00  1.270e+00  -3.152 0.001623 ** 
## hospital                  4.157e+00  6.961e-01   5.971 2.39e-09 ***
## doctors                   3.432e+00  6.210e-01   5.527 3.30e-08 ***
## theatre                   3.271e+00  7.462e-01   4.384 1.17e-05 ***
## university                4.529e-01  8.946e-01   0.506 0.612646    
## clock                     6.010e+01  5.375e+00  11.181  < 2e-16 ***
## parking_entrance         -3.560e-02  1.985e+00  -0.018 0.985692    
## police                   -5.186e+00  8.438e-01  -6.146 8.07e-10 ***
## cultural_center           7.626e+01  5.024e+00  15.178  < 2e-16 ***
## stripclub                -4.049e-01  1.751e+00  -0.231 0.817158    
## marketplace               8.374e+00  1.131e+00   7.402 1.39e-13 ***
## dry_cleaner              -5.154e+00  1.886e+00  -2.733 0.006287 ** 
## bicycle_repair_station   -5.909e+00  1.051e+00  -5.623 1.90e-08 ***
## office                    3.275e+01  3.021e+00  10.843  < 2e-16 ***
## arts_centre               4.264e+00  1.228e+00   3.472 0.000518 ***
## library                  -2.608e+00  1.112e+00  -2.346 0.018967 *  
## studio                   -1.997e+00  2.662e+00  -0.750 0.453304    
## strip_club                9.143e+00  2.763e+00   3.309 0.000937 ***
## tourist                   8.367e+00  9.179e-01   9.116  < 2e-16 ***
## veterinary                4.362e+00  2.243e+00   1.944 0.051853 .  
```

```
## community_centre          1.972e+01  2.862e+00   6.889 5.78e-12 ***
## compressed_air            1.314e+01  1.752e+00   7.501 6.54e-14 ***
## tutor                     4.094e+00  1.942e+00   2.108 0.035009 *
## clinic                    1.322e+00  1.193e+00   1.107 0.268097
## dentist                  -5.145e+00  1.483e+00  -3.471 0.000520 ***
## bench                    -1.302e+00  1.620e-01  -8.039 9.47e-16 ***
## cinema                   -1.861e+00  2.693e+00  -0.691 0.489546
## college                  -1.752e+01  2.518e+00  -6.958 3.54e-12 ***
## parking_exit              8.442e+00  2.912e+00   2.899 0.003746 **
## bar.restaurant            8.397e+00  4.680e+00   1.794 0.072791 .
## car_rental               -3.744e+00  2.588e+00  -1.447 0.147931
## coworking                -8.957e+00  4.449e+00  -2.013 0.044110 *
## shelter                  -1.165e+01  1.144e+00 -10.179  < 2e-16 ***
## bureau_de_change          1.893e+01  3.236e+00   5.850 4.98e-09 ***
## food_cart                -2.863e+01  3.871e+00  -7.395 1.46e-13 ***
## school..historic.        -1.990e-01  2.722e+00  -0.073 0.941701
## border_control           -5.660e+00  4.644e+00  -1.219 0.222966
## check_cashing            -4.625e+00  3.739e+00  -1.237 0.216100
## nail_salon                2.131e+01  5.537e+00   3.849 0.000119 ***
## storage                          NA         NA      NA       NA
## tax                      -6.132e-01  8.252e+00  -0.074 0.940771
## catering                  1.246e+01  3.097e+00   4.023 5.75e-05 ***
## dojo                             NA         NA      NA       NA
## tax_service                      NA         NA      NA       NA
## bus_station               2.360e+01  2.476e+00   9.531  < 2e-16 ***
## hospital..historic.              NA         NA      NA       NA
## toilets                  -7.618e+00  1.554e+00  -4.901 9.58e-07 ***
## marker                    3.165e+00  2.193e+00   1.444 0.148854
## social_facility          -1.076e+01  3.731e+00  -2.883 0.003939 **
## telephone                        NA         NA      NA       NA
## taxi                     -1.170e+01  2.138e+00  -5.474 4.45e-08 ***
## building                         NA         NA      NA       NA
## gym                              NA         NA      NA       NA
## emergency_phone           1.178e+01  3.381e+00   3.483 0.000497 ***
## courthouse                       NA         NA      NA       NA
## fitness_center            7.188e+00  3.403e+00   2.112 0.034674 *
## townhall                 -1.307e+01  3.223e+00  -4.055 5.02e-05 ***
## car_wash                         NA         NA      NA       NA
## ev_charging                      NA         NA      NA       NA
## recycling.waste_basket           NA         NA      NA       NA
## sign                             NA         NA      NA       NA
## charging_station                 NA         NA      NA       NA
## photography                      NA         NA      NA       NA
## picnic_table                     NA         NA      NA       NA
## nursing_home                     NA         NA      NA       NA
## traffic_signals           1.133e-01  1.922e-02   5.895 3.80e-09 ***
## crossing                  1.173e-01  1.525e-02   7.693 1.49e-14 ***
## motorway_junction        -6.763e-01  2.339e-01  -2.892 0.003834 **
## bus_stop                  5.529e-01  1.911e-01   2.894 0.003810 **
## speed_camera              1.047e+01  1.889e+00   5.544 2.99e-08 ***
## service                          NA         NA      NA       NA
## stop                     -8.593e-02  1.069e-01  -0.804 0.421534
## turning_circle           -2.364e+00  6.947e-01  -3.403 0.000668 ***
## elevator                 -6.322e-02  4.563e+00  -0.014 0.988946
```

```
## traffic_signals.bus_stop -7.627e+00  2.332e+00  -3.270 0.001075 **
## mini_roundabout          -1.197e+01  1.697e+00  -7.052 1.81e-12 ***
## footway                  -8.460e+00  2.181e+00  -3.880 0.000105 ***
## street_lamp               7.250e+00  1.811e+00   4.005 6.23e-05 ***
## turning_loop                     NA         NA      NA       NA
## hotel                     2.509e+00  2.817e-01   8.907  < 2e-16 ***
## artwork                   4.910e-02  1.385e-01   0.355 0.722955
## information              -6.383e+00  8.432e-01  -7.570 3.87e-14 ***
## museum                   -2.024e+00  7.484e-01  -2.705 0.006842 **
## sculpture                -1.059e+00  6.480e-01  -1.634 0.102320
## hostel                   -3.427e+01  4.472e+00  -7.662 1.90e-14 ***
## picnic_site               8.514e-01  8.184e-01   1.040 0.298183
## tour_guide               -6.260e+00  4.253e+00  -1.472 0.141051
## attraction               -3.065e+00  5.734e-01  -5.344 9.17e-08 ***
## landmark                 -1.204e+01  2.057e+00  -5.853 4.91e-09 ***
## motel                    -9.610e+00  2.864e+00  -3.355 0.000794 ***
## guest_house              -1.725e+01  3.620e+00  -4.765 1.90e-06 ***
## gallery                  -2.234e-01  1.108e+00  -0.202 0.840265
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.97 on 23269 degrees of freedom
## Multiple R-squared:  0.5252, Adjusted R-squared:  0.5227
## F-statistic: 214.5 on 120 and 23269 DF,  p-value: < 2.2e-16
```

```
# hmm, we have some weirdness in there, some stations features don't exist
# around any of our stations, e.g. 'turning_loop'
table(data_to_model$turning_loop)
```

```
##
##     0
## 23390
```

```
# lets remove those using the handly 'colSums' and 'which' functions
colSums(data_to_model[ , 15:143])
```

```
##               fast_food                 parking              restaurant
##                   41155                   15531                  112007
##             convenience             post_office         bicycle_parking
##                     684                    5156                    6528
##          drinking_water               recycling            waste_basket
##                    7462                     968                    7036
##          waste_disposal                    cafe       currency_exchange
##                       0                   42066                     150
##                fountain               ice_cream                optician
##                    2387                       0                       0
##                pharmacy           tanning_salon             car_sharing
##                    8493                       0                     564
##                 alcohol                    bank                     bar
##                     294                   37061                   18419
##                    club                 embassy              food_court
##                     677                    3062                    1003
##              government  internal_kindergarten             kindergarten
```

```
##                   551                   381                  5492
##       place_of_worship              post_box                   pub
##                 21231                 10382                 17565
##        vending_machine                  fuel             grave_yard
##                   381                  1695                  1154
##        public_building                school          fire_station
##                  2560                 17131                  1532
##             nightclub                   atm              hospital
##                  2472                  2122                   779
##               doctors               theatre            university
##                  1955                  2777                   721
##                 clock      parking_entrance                police
##                   238                  1167                  1232
##        cultural_center             stripclub           marketplace
##                   297                   295                  1195
##            dry_cleaner bicycle_repair_station                office
##                   354                   876                   533
##            arts_centre               library                studio
##                  2219                  1390                   282
##             strip_club               tourist            veterinary
##                   382                  2719                   741
##       community_centre        compressed_air                 tutor
##                   300                   302                   300
##                clinic               dentist                 bench
##                   970                   420                  6830
##                cinema               college          parking_exit
##                   823                   297                   143
##         bar.restaurant            car_rental             coworking
##                   148                   973                   144
##               shelter       bureau_de_change             food_cart
##                   671                   232                   320
##      school..historic.        border_control         check_cashing
##                   148                   299                   434
##             nail_salon               storage                   tax
##                   145                   145                    71
##               catering                  dojo           tax_service
##                   241                   147                   147
##            bus_station   hospital..historic.               toilets
##                   150                     0                   613
##                 marker       social_facility             telephone
##                   108                   139                   139
##                   taxi              building                   gym
##                   324                     0                     0
##        emergency_phone            courthouse        fitness_center
##                   296                     0                    98
##               townhall              car_wash           ev_charging
##                    84                     0                     0
## recycling.waste_basket                  sign      charging_station
##                     0                     0                     0
##            photography           picnic_table          nursing_home
##                     0                     0                     0
##        traffic_signals              crossing      motorway_junction
##                440230                514335                  4789
##               bus_stop          speed_camera               service
```

```
##                    33156                     148                       0
##                     stop          turning_circle                elevator
##                    25580                    1276                     150
## traffic_signals.bus_stop          mini_roundabout                  footway
##                      144                     335                      98
##              street_lamp            turning_loop                   hotel
##                      150                       0                   15907
##                  artwork             information                  museum
##                    12172                    2117                    2762
##                sculpture                  hostel             picnic_site
##                      564                     533                     569
##               tour_guide              attraction                landmark
##                      239                    2423                     268
##                    motel             guest_house                 gallery
##                       73                      56                    1187
```

```r
# we want to know 'which' columns have a sum of 0
columns_to_remove = names(which(colSums(data_to_model[ , 15:143]) == 0))

# now combine that with filter to remove those from our data
data_to_model = data_to_model[ , !(names(data_to_model) %in% columns_to_remove)]

# try the model again
model = lm(rentals ~ ., data = data_to_model)
summary(model)
```

```
##
## Call:
## lm(formula = rentals ~ ., data = data_to_model)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.933 -11.365  -0.838   8.980 222.391
##
## Coefficients: (5 not defined because of singularities)
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -1.737e+01  1.618e+00 -10.740  < 2e-16 ***
## cust_typeRegistered      2.755e+01  2.854e-01  96.523  < 2e-16 ***
## cust_typeSubscriber      1.603e+01  4.039e-01  39.699  < 2e-16 ***
## weekday                  6.292e-01  6.487e-02   9.700  < 2e-16 ***
## season_code             -5.139e-01  1.404e-01  -3.662 0.000251 ***
## is_holiday              -3.132e+00  4.912e-01  -6.376 1.85e-10 ***
## is_work_day              2.061e+00  2.851e-01   7.231 4.94e-13 ***
## weather_code            -3.945e+00  4.162e-01  -9.479  < 2e-16 ***
## duration                 8.699e-04  4.891e-04   1.779 0.075313 .
## temp                    -1.483e+02  1.087e+01 -13.637  < 2e-16 ***
## subjective_temp          1.700e+02  1.262e+01  13.464  < 2e-16 ***
## humidity                -7.251e-01  2.299e+00  -0.315 0.752438
## windspeed               -1.540e+01  2.692e+00  -5.719 1.08e-08 ***
## no_bikes                 8.066e-01  3.916e-02  20.597  < 2e-16 ***
## no_empty_docks           6.027e-01  3.984e-02  15.129  < 2e-16 ***
## fast_food                4.224e-01  1.623e-01   2.602 0.009271 **
## parking                 -1.321e+00  3.253e-01  -4.062 4.89e-05 ***
## restaurant               1.273e-01  6.479e-02   1.965 0.049453 *
```

```
## convenience            -2.393e+01  2.985e+00  -8.016 1.14e-15 ***
## post_office             7.823e-02  6.808e-01   0.115 0.908520
## bicycle_parking        -2.040e+00  5.089e-01  -4.008 6.14e-05 ***
## drinking_water          1.715e+00  3.819e-01   4.490 7.15e-06 ***
## recycling              -6.329e+00  1.719e+00  -3.681 0.000233 ***
## waste_basket            1.243e+00  2.504e-01   4.963 6.98e-07 ***
## cafe                    6.801e-01  2.560e-01   2.656 0.007902 **
## currency_exchange       3.960e+01  2.482e+00  15.957  < 2e-16 ***
## fountain                3.233e+00  6.346e-01   5.095 3.52e-07 ***
## pharmacy                1.131e+00  6.380e-01   1.772 0.076344 .
## car_sharing             2.301e+00  1.083e+00   2.124 0.033657 *
## alcohol                 3.924e+00  2.349e+00   1.670 0.094870 .
## bank                   -2.075e+00  1.681e-01 -12.344  < 2e-16 ***
## bar                     1.650e+00  1.466e-01  11.252  < 2e-16 ***
## club                   -1.294e+01  3.079e+00  -4.203 2.64e-05 ***
## embassy                 8.422e+00  7.531e-01  11.182  < 2e-16 ***
## food_court             -1.314e+01  1.974e+00  -6.656 2.87e-11 ***
## government              6.923e+00  3.857e+00   1.795 0.072650 .
## internal_kindergarten   2.657e+01  4.409e+00   6.026 1.70e-09 ***
## kindergarten           -1.319e-02  4.306e-01  -0.031 0.975564
## place_of_worship        1.489e+00  1.405e-01  10.594  < 2e-16 ***
## post_box               -1.685e+00  5.213e-01  -3.232 0.001231 **
## pub                     4.257e+00  2.776e-01  15.338  < 2e-16 ***
## vending_machine                NA         NA      NA       NA
## fuel                    1.432e+00  8.514e-01   1.682 0.092657 .
## grave_yard              2.614e+00  7.687e-01   3.401 0.000672 ***
## public_building         2.954e+00  6.115e-01   4.830 1.37e-06 ***
## school                  2.172e+00  1.732e-01  12.542  < 2e-16 ***
## fire_station           -2.859e-01  8.600e-01  -0.332 0.739528
## nightclub              -1.673e+01  1.266e+00 -13.215  < 2e-16 ***
## atm                    -4.002e+00  1.270e+00  -3.152 0.001623 **
## hospital                4.157e+00  6.961e-01   5.971 2.39e-09 ***
## doctors                 3.432e+00  6.210e-01   5.527 3.30e-08 ***
## theatre                 3.271e+00  7.462e-01   4.384 1.17e-05 ***
## university              4.529e-01  8.946e-01   0.506 0.612646
## clock                   6.010e+01  5.375e+00  11.181  < 2e-16 ***
## parking_entrance       -3.560e-02  1.985e+00  -0.018 0.985692
## police                 -5.186e+00  8.438e-01  -6.146 8.07e-10 ***
## cultural_center         7.626e+01  5.024e+00  15.178  < 2e-16 ***
## stripclub              -4.049e-01  1.751e+00  -0.231 0.817158
## marketplace             8.374e+00  1.131e+00   7.402 1.39e-13 ***
## dry_cleaner            -5.154e+00  1.886e+00  -2.733 0.006287 **
## bicycle_repair_station -5.909e+00  1.051e+00  -5.623 1.90e-08 ***
## office                  3.275e+01  3.021e+00  10.843  < 2e-16 ***
## arts_centre             4.264e+00  1.228e+00   3.472 0.000518 ***
## library                -2.608e+00  1.112e+00  -2.346 0.018967 *
## studio                 -1.997e+00  2.662e+00  -0.750 0.453304
## strip_club              9.143e+00  2.763e+00   3.309 0.000937 ***
## tourist                 8.367e+00  9.179e-01   9.116  < 2e-16 ***
## veterinary              4.362e+00  2.243e+00   1.944 0.051853 .
## community_centre        1.972e+01  2.862e+00   6.889 5.78e-12 ***
## compressed_air          1.314e+01  1.752e+00   7.501 6.54e-14 ***
## tutor                   4.094e+00  1.942e+00   2.108 0.035009 *
## clinic                  1.322e+00  1.193e+00   1.107 0.268097
```

```
## dentist                 -5.145e+00  1.483e+00  -3.471 0.000520 ***
## bench                   -1.302e+00  1.620e-01  -8.039 9.47e-16 ***
## cinema                  -1.861e+00  2.693e+00  -0.691 0.489546
## college                 -1.752e+01  2.518e+00  -6.958 3.54e-12 ***
## parking_exit             8.442e+00  2.912e+00   2.899 0.003746 **
## bar.restaurant           8.397e+00  4.680e+00   1.794 0.072791 .
## car_rental              -3.744e+00  2.588e+00  -1.447 0.147931
## coworking               -8.957e+00  4.449e+00  -2.013 0.044110 *
## shelter                 -1.165e+01  1.144e+00 -10.179  < 2e-16 ***
## bureau_de_change         1.893e+01  3.236e+00   5.850 4.98e-09 ***
## food_cart               -2.863e+01  3.871e+00  -7.395 1.46e-13 ***
## school..historic.       -1.990e-01  2.722e+00  -0.073 0.941701
## border_control          -5.660e+00  4.644e+00  -1.219 0.222966
## check_cashing           -4.625e+00  3.739e+00  -1.237 0.216100
## nail_salon               2.131e+01  5.537e+00   3.849 0.000119 ***
## storage                         NA         NA      NA       NA
## tax                     -6.132e-01  8.252e+00  -0.074 0.940771
## catering                 1.246e+01  3.097e+00   4.023 5.75e-05 ***
## dojo                            NA         NA      NA       NA
## tax_service                     NA         NA      NA       NA
## bus_station              2.360e+01  2.476e+00   9.531  < 2e-16 ***
## toilets                 -7.618e+00  1.554e+00  -4.901 9.58e-07 ***
## marker                   3.165e+00  2.193e+00   1.444 0.148854
## social_facility         -1.076e+01  3.731e+00  -2.883 0.003939 **
## telephone                       NA         NA      NA       NA
## taxi                    -1.170e+01  2.138e+00  -5.474 4.45e-08 ***
## emergency_phone          1.178e+01  3.381e+00   3.483 0.000497 ***
## fitness_center           7.188e+00  3.403e+00   2.112 0.034674 *
## townhall                -1.307e+01  3.223e+00  -4.055 5.02e-05 ***
## traffic_signals          1.133e-01  1.922e-02   5.895 3.80e-09 ***
## crossing                 1.173e-01  1.525e-02   7.693 1.49e-14 ***
## motorway_junction       -6.763e-01  2.339e-01  -2.892 0.003834 **
## bus_stop                 5.529e-01  1.911e-01   2.894 0.003810 **
## speed_camera             1.047e+01  1.889e+00   5.544 2.99e-08 ***
## stop                    -8.593e-02  1.069e-01  -0.804 0.421534
## turning_circle          -2.364e+00  6.947e-01  -3.403 0.000668 ***
## elevator                -6.322e-02  4.563e+00  -0.014 0.988946
## traffic_signals.bus_stop -7.627e+00  2.332e+00  -3.270 0.001075 **
## mini_roundabout         -1.197e+01  1.697e+00  -7.052 1.81e-12 ***
## footway                 -8.460e+00  2.181e+00  -3.880 0.000105 ***
## street_lamp              7.250e+00  1.811e+00   4.005 6.23e-05 ***
## hotel                    2.509e+00  2.817e-01   8.907  < 2e-16 ***
## artwork                  4.910e-02  1.385e-01   0.355 0.722955
## information             -6.383e+00  8.432e-01  -7.570 3.87e-14 ***
## museum                  -2.024e+00  7.484e-01  -2.705 0.006842 **
## sculpture               -1.059e+00  6.480e-01  -1.634 0.102320
## hostel                  -3.427e+01  4.472e+00  -7.662 1.90e-14 ***
## picnic_site              8.514e-01  8.184e-01   1.040 0.298183
## tour_guide              -6.260e+00  4.253e+00  -1.472 0.141051
## attraction              -3.065e+00  5.734e-01  -5.344 9.17e-08 ***
## landmark                -1.204e+01  2.057e+00  -5.853 4.91e-09 ***
## motel                   -9.610e+00  2.864e+00  -3.355 0.000794 ***
## guest_house             -1.725e+01  3.620e+00  -4.765 1.90e-06 ***
## gallery                 -2.234e-01  1.108e+00  -0.202 0.840265
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.97 on 23269 degrees of freedom
## Multiple R-squared:  0.5252, Adjusted R-squared:  0.5227
## F-statistic: 214.5 on 120 and 23269 DF,  p-value: < 2.2e-16
```

```r
# definintely better, but we still have some weird NAs, lets troubleshoot those
table(data_to_model$vending_machine)
```

```
##
##     0     1
## 23009   381
```

```r
table(data_to_model$storage)
```

```
##
##     0     1
## 23245   145
```

```r
table(data_to_model$dojo)
```

```
##
##     0     1
## 23243   147
```

```r
table(data_to_model$tax_service)
```

```
##
##     0     1
## 23243   147
```

```r
table(data_to_model$telephone)
```

```
##
##     0     1
## 23251   139
```

```r
# all the landmarks have at most 1 in the area, so there are not enough
# observations for least square to fit the model.
# these variables won't be helpful in prediction, so lets remove them.

data_to_model =
  data_to_model %>%
  select(
    -vending_machine,
    -storage,
    -dojo,
    -tax_service,
    -telephone)
```

```r
# try the model again
model = lm(rentals ~ ., data = data_to_model)
summary(model)
```

```
##
## Call:
## lm(formula = rentals ~ ., data = data_to_model)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.933 -11.365  -0.838   8.980 222.391
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -1.737e+01  1.618e+00 -10.740  < 2e-16 ***
## cust_typeRegistered       2.755e+01  2.854e-01  96.523  < 2e-16 ***
## cust_typeSubscriber       1.603e+01  4.039e-01  39.699  < 2e-16 ***
## weekday                   6.292e-01  6.487e-02   9.700  < 2e-16 ***
## season_code              -5.139e-01  1.404e-01  -3.662 0.000251 ***
## is_holiday               -3.132e+00  4.912e-01  -6.376 1.85e-10 ***
## is_work_day               2.061e+00  2.851e-01   7.231 4.94e-13 ***
## weather_code             -3.945e+00  4.162e-01  -9.479  < 2e-16 ***
## duration                  8.699e-04  4.891e-04   1.779 0.075313 .
## temp                     -1.483e+02  1.087e+01 -13.637  < 2e-16 ***
## subjective_temp           1.700e+02  1.262e+01  13.464  < 2e-16 ***
## humidity                 -7.251e-01  2.299e+00  -0.315 0.752438
## windspeed                -1.540e+01  2.692e+00  -5.719 1.08e-08 ***
## no_bikes                  8.066e-01  3.916e-02  20.597  < 2e-16 ***
## no_empty_docks            6.027e-01  3.984e-02  15.129  < 2e-16 ***
## fast_food                 4.224e-01  1.623e-01   2.602 0.009271 **
## parking                  -1.321e+00  3.253e-01  -4.062 4.89e-05 ***
## restaurant                1.273e-01  6.479e-02   1.965 0.049453 *
## convenience              -2.393e+01  2.985e+00  -8.016 1.14e-15 ***
## post_office               7.823e-02  6.808e-01   0.115 0.908520
## bicycle_parking          -2.040e+00  5.089e-01  -4.008 6.14e-05 ***
## drinking_water            1.715e+00  3.819e-01   4.490 7.15e-06 ***
## recycling                -6.329e+00  1.719e+00  -3.681 0.000233 ***
## waste_basket              1.243e+00  2.504e-01   4.963 6.98e-07 ***
## cafe                      6.801e-01  2.560e-01   2.656 0.007902 **
## currency_exchange         3.960e+01  2.482e+00  15.957  < 2e-16 ***
## fountain                  3.233e+00  6.346e-01   5.095 3.52e-07 ***
## pharmacy                  1.131e+00  6.380e-01   1.772 0.076344 .
## car_sharing               2.301e+00  1.083e+00   2.124 0.033657 *
## alcohol                   3.924e+00  2.349e+00   1.670 0.094870 .
## bank                     -2.075e+00  1.681e-01 -12.344  < 2e-16 ***
## bar                       1.650e+00  1.466e-01  11.252  < 2e-16 ***
## club                     -1.294e+01  3.079e+00  -4.203 2.64e-05 ***
## embassy                   8.422e+00  7.531e-01  11.182  < 2e-16 ***
## food_court               -1.314e+01  1.974e+00  -6.656 2.87e-11 ***
## government                6.923e+00  3.857e+00   1.795 0.072650 .
## internal_kindergarten     2.657e+01  4.409e+00   6.026 1.70e-09 ***
## kindergarten             -1.319e-02  4.306e-01  -0.031 0.975564
## place_of_worship          1.489e+00  1.405e-01  10.594  < 2e-16 ***
```

```
## post_box                -1.685e+00  5.213e-01  -3.232 0.001231 **
## pub                      4.257e+00  2.776e-01  15.338  < 2e-16 ***
## fuel                     1.432e+00  8.514e-01   1.682 0.092657 .
## grave_yard               2.614e+00  7.687e-01   3.401 0.000672 ***
## public_building          2.954e+00  6.115e-01   4.830 1.37e-06 ***
## school                   2.172e+00  1.732e-01  12.542  < 2e-16 ***
## fire_station            -2.859e-01  8.600e-01  -0.332 0.739528
## nightclub               -1.673e+01  1.266e+00 -13.215  < 2e-16 ***
## atm                     -4.002e+00  1.270e+00  -3.152 0.001623 **
## hospital                 4.157e+00  6.961e-01   5.971 2.39e-09 ***
## doctors                  3.432e+00  6.210e-01   5.527 3.30e-08 ***
## theatre                  3.271e+00  7.462e-01   4.384 1.17e-05 ***
## university               4.529e-01  8.946e-01   0.506 0.612646
## clock                    6.010e+01  5.375e+00  11.181  < 2e-16 ***
## parking_entrance        -3.560e-02  1.985e+00  -0.018 0.985692
## police                  -5.186e+00  8.438e-01  -6.146 8.07e-10 ***
## cultural_center          7.626e+01  5.024e+00  15.178  < 2e-16 ***
## stripclub               -4.049e-01  1.751e+00  -0.231 0.817158
## marketplace              8.374e+00  1.131e+00   7.402 1.39e-13 ***
## dry_cleaner             -5.154e+00  1.886e+00  -2.733 0.006287 **
## bicycle_repair_station  -5.909e+00  1.051e+00  -5.623 1.90e-08 ***
## office                   3.275e+01  3.021e+00  10.843  < 2e-16 ***
## arts_centre              4.264e+00  1.228e+00   3.472 0.000518 ***
## library                 -2.608e+00  1.112e+00  -2.346 0.018967 *
## studio                  -1.997e+00  2.662e+00  -0.750 0.453304
## strip_club               9.143e+00  2.763e+00   3.309 0.000937 ***
## tourist                  8.367e+00  9.179e-01   9.116  < 2e-16 ***
## veterinary               4.362e+00  2.243e+00   1.944 0.051853 .
## community_centre         1.972e+01  2.862e+00   6.889 5.78e-12 ***
## compressed_air           1.314e+01  1.752e+00   7.501 6.54e-14 ***
## tutor                    4.094e+00  1.942e+00   2.108 0.035009 *
## clinic                   1.322e+00  1.193e+00   1.107 0.268097
## dentist                 -5.145e+00  1.483e+00  -3.471 0.000520 ***
## bench                   -1.302e+00  1.620e-01  -8.039 9.47e-16 ***
## cinema                  -1.861e+00  2.693e+00  -0.691 0.489546
## college                 -1.752e+01  2.518e+00  -6.958 3.54e-12 ***
## parking_exit             8.442e+00  2.912e+00   2.899 0.003746 **
## bar.restaurant           8.397e+00  4.680e+00   1.794 0.072791 .
## car_rental              -3.744e+00  2.588e+00  -1.447 0.147931
## coworking               -8.957e+00  4.449e+00  -2.013 0.044110 *
## shelter                 -1.165e+01  1.144e+00 -10.179  < 2e-16 ***
## bureau_de_change         1.893e+01  3.236e+00   5.850 4.98e-09 ***
## food_cart               -2.863e+01  3.871e+00  -7.395 1.46e-13 ***
## school..historic.       -1.990e-01  2.722e+00  -0.073 0.941701
## border_control          -5.660e+00  4.644e+00  -1.219 0.222966
## check_cashing           -4.625e+00  3.739e+00  -1.237 0.216100
## nail_salon               2.131e+01  5.537e+00   3.849 0.000119 ***
## tax                     -6.132e-01  8.252e+00  -0.074 0.940771
## catering                 1.246e+01  3.097e+00   4.023 5.75e-05 ***
## bus_station              2.360e+01  2.476e+00   9.531  < 2e-16 ***
## toilets                 -7.618e+00  1.554e+00  -4.901 9.58e-07 ***
## marker                   3.165e+00  2.193e+00   1.444 0.148854
## social_facility         -1.076e+01  3.731e+00  -2.883 0.003939 **
## taxi                    -1.170e+01  2.138e+00  -5.474 4.45e-08 ***
```

```
## emergency_phone           1.178e+01  3.381e+00   3.483 0.000497 ***
## fitness_center            7.188e+00  3.403e+00   2.112 0.034674 *
## townhall                 -1.307e+01  3.223e+00  -4.055 5.02e-05 ***
## traffic_signals           1.133e-01  1.922e-02   5.895 3.80e-09 ***
## crossing                  1.173e-01  1.525e-02   7.693 1.49e-14 ***
## motorway_junction        -6.763e-01  2.339e-01  -2.892 0.003834 **
## bus_stop                  5.529e-01  1.911e-01   2.894 0.003810 **
## speed_camera              1.047e+01  1.889e+00   5.544 2.99e-08 ***
## stop                     -8.593e-02  1.069e-01  -0.804 0.421534
## turning_circle           -2.364e+00  6.947e-01  -3.403 0.000668 ***
## elevator                 -6.322e-02  4.563e+00  -0.014 0.988946
## traffic_signals.bus_stop -7.627e+00  2.332e+00  -3.270 0.001075 **
## mini_roundabout          -1.197e+01  1.697e+00  -7.052 1.81e-12 ***
## footway                  -8.460e+00  2.181e+00  -3.880 0.000105 ***
## street_lamp               7.250e+00  1.811e+00   4.005 6.23e-05 ***
## hotel                     2.509e+00  2.817e-01   8.907  < 2e-16 ***
## artwork                   4.910e-02  1.385e-01   0.355 0.722955
## information              -6.383e+00  8.432e-01  -7.570 3.87e-14 ***
## museum                   -2.024e+00  7.484e-01  -2.705 0.006842 **
## sculpture                -1.059e+00  6.480e-01  -1.634 0.102320
## hostel                   -3.427e+01  4.472e+00  -7.662 1.90e-14 ***
## picnic_site               8.514e-01  8.184e-01   1.040 0.298183
## tour_guide               -6.260e+00  4.253e+00  -1.472 0.141051
## attraction               -3.065e+00  5.734e-01  -5.344 9.17e-08 ***
## landmark                 -1.204e+01  2.057e+00  -5.853 4.91e-09 ***
## motel                    -9.610e+00  2.864e+00  -3.355 0.000794 ***
## guest_house              -1.725e+01  3.620e+00  -4.765 1.90e-06 ***
## gallery                  -2.234e-01  1.108e+00  -0.202 0.840265
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.97 on 23269 degrees of freedom
## Multiple R-squared:  0.5252, Adjusted R-squared:  0.5227
## F-statistic: 214.5 on 120 and 23269 DF,  p-value: < 2.2e-16
```

```r
# one last modification. our categorical variables are being treated like
# they're continuous. lets create some factors
data_to_model$weekday = factor(data_to_model$weekday,
                               labels = 0:6,
                               levels = 0:6)
data_to_model$season_code = factor(data_to_model$season_code)
data_to_model$is_holiday = factor(data_to_model$is_holiday)
data_to_model$is_work_day = factor(data_to_model$is_work_day)
data_to_model$weather_code = factor(data_to_model$weather_code)

# try the model again
model = lm(rentals ~ ., data = data_to_model)
summary(model)
```

```
##
## Call:
## lm(formula = rentals ~ ., data = data_to_model)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -68.437 -11.258  -0.955   8.856 216.911
##
## Coefficients: (1 not defined because of singularities)
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -2.735e+01  1.835e+00 -14.908  < 2e-16 ***
## cust_typeRegistered       2.732e+01  2.849e-01  95.899  < 2e-16 ***
## cust_typeSubscriber       1.615e+01  4.026e-01  40.109  < 2e-16 ***
## weekday1                  2.304e+00  4.708e-01   4.893 1.00e-06 ***
## weekday2                  1.862e+00  4.875e-01   3.820 0.000134 ***
## weekday3                  2.734e+00  4.976e-01   5.495 3.95e-08 ***
## weekday4                  2.244e+00  4.886e-01   4.593 4.40e-06 ***
## weekday5                  4.683e+00  5.130e-01   9.129  < 2e-16 ***
## weekday6                  3.474e+00  4.690e-01   7.408 1.33e-13 ***
## season_code2             -7.723e+00  6.074e-01 -12.715  < 2e-16 ***
## season_code3             -1.244e+01  8.393e-01 -14.820  < 2e-16 ***
## season_code4             -4.512e+00  4.975e-01  -9.069  < 2e-16 ***
## is_holiday1              -5.004e+00  4.779e-01 -10.472  < 2e-16 ***
## is_work_day1                     NA         NA      NA       NA
## weather_code2            -2.382e+00  4.441e-01  -5.363 8.28e-08 ***
## weather_code3            -9.405e+00  9.631e-01  -9.766  < 2e-16 ***
## duration                  8.655e-04  4.863e-04   1.780 0.075122 .
## temp                     -9.362e+01  1.273e+01  -7.353 2.00e-13 ***
## subjective_temp           1.369e+02  1.437e+01   9.525  < 2e-16 ***
## humidity                 -2.456e+00  2.304e+00  -1.066 0.286593
## windspeed                -1.303e+01  2.845e+00  -4.581 4.66e-06 ***
## no_bikes                  8.103e-01  3.893e-02  20.814  < 2e-16 ***
## no_empty_docks            6.043e-01  3.961e-02  15.257  < 2e-16 ***
## fast_food                 4.193e-01  1.614e-01   2.598 0.009379 **
## parking                  -1.338e+00  3.234e-01  -4.137 3.53e-05 ***
## restaurant                1.306e-01  6.441e-02   2.027 0.042667 *
## convenience              -2.416e+01  2.967e+00  -8.143 4.05e-16 ***
## post_office               2.378e-01  6.769e-01   0.351 0.725334
## bicycle_parking          -2.052e+00  5.059e-01  -4.056 5.02e-05 ***
## drinking_water            1.743e+00  3.797e-01   4.591 4.43e-06 ***
## recycling                -5.961e+00  1.710e+00  -3.487 0.000490 ***
## waste_basket              1.201e+00  2.490e-01   4.825 1.41e-06 ***
## cafe                      7.180e-01  2.546e-01   2.821 0.004798 **
## currency_exchange         4.004e+01  2.468e+00  16.224  < 2e-16 ***
## fountain                  3.241e+00  6.308e-01   5.138 2.80e-07 ***
## pharmacy                  1.102e+00  6.342e-01   1.738 0.082190 .
## car_sharing               2.585e+00  1.077e+00   2.399 0.016426 *
## alcohol                   3.801e+00  2.335e+00   1.627 0.103650
## bank                     -2.106e+00  1.672e-01 -12.600  < 2e-16 ***
## bar                       1.639e+00  1.458e-01  11.247  < 2e-16 ***
## club                     -1.305e+01  3.061e+00  -4.264 2.01e-05 ***
## embassy                   8.440e+00  7.487e-01  11.273  < 2e-16 ***
## food_court               -1.288e+01  1.962e+00  -6.563 5.37e-11 ***
## government                6.668e+00  3.834e+00   1.739 0.082017 .
## internal_kindergarten     2.661e+01  4.383e+00   6.071 1.29e-09 ***
## kindergarten              3.185e-02  4.280e-01   0.074 0.940680
## place_of_worship          1.496e+00  1.397e-01  10.707  < 2e-16 ***
## post_box                 -1.668e+00  5.183e-01  -3.218 0.001293 **
## pub                       4.263e+00  2.759e-01  15.451  < 2e-16 ***
```

```
## fuel                     1.376e+00  8.465e-01   1.625 0.104155    
## grave_yard               2.618e+00  7.642e-01   3.426 0.000614 ***
## public_building          3.007e+00  6.079e-01   4.947 7.61e-07 ***
## school                   2.194e+00  1.722e-01  12.745  < 2e-16 ***
## fire_station            -4.472e-01  8.551e-01  -0.523 0.601037    
## nightclub               -1.675e+01  1.258e+00 -13.310  < 2e-16 ***
## atm                     -4.184e+00  1.262e+00  -3.315 0.000918 ***
## hospital                 4.134e+00  6.920e-01   5.974 2.34e-09 ***
## doctors                  3.499e+00  6.174e-01   5.667 1.47e-08 ***
## theatre                  3.360e+00  7.419e-01   4.529 5.95e-06 ***
## university               4.302e-01  8.893e-01   0.484 0.628569    
## clock                    5.920e+01  5.345e+00  11.075  < 2e-16 ***
## parking_entrance         2.608e-01  1.974e+00   0.132 0.894903    
## police                  -5.191e+00  8.388e-01  -6.189 6.16e-10 ***
## cultural_center          7.667e+01  4.995e+00  15.348  < 2e-16 ***
## stripclub               -4.043e-01  1.741e+00  -0.232 0.816356    
## marketplace              8.263e+00  1.125e+00   7.347 2.10e-13 ***
## dry_cleaner             -4.864e+00  1.875e+00  -2.594 0.009493 ** 
## bicycle_repair_station  -5.854e+00  1.045e+00  -5.603 2.13e-08 ***
## office                   3.302e+01  3.003e+00  10.996  < 2e-16 ***
## arts_centre              4.082e+00  1.221e+00   3.343 0.000830 ***
## library                 -2.727e+00  1.105e+00  -2.468 0.013599 *  
## studio                  -1.853e+00  2.647e+00  -0.700 0.483849    
## strip_club               9.008e+00  2.747e+00   3.279 0.001043 ** 
## tourist                  8.292e+00  9.126e-01   9.086  < 2e-16 ***
## veterinary               4.297e+00  2.230e+00   1.927 0.053990 .  
## community_centre         1.978e+01  2.845e+00   6.953 3.67e-12 ***
## compressed_air           1.316e+01  1.742e+00   7.557 4.28e-14 ***
## tutor                    3.795e+00  1.931e+00   1.966 0.049355 *  
## clinic                   1.310e+00  1.186e+00   1.104 0.269568    
## dentist                 -5.294e+00  1.474e+00  -3.592 0.000329 ***
## bench                   -1.307e+00  1.611e-01  -8.117 5.00e-16 ***
## cinema                  -1.815e+00  2.678e+00  -0.678 0.497841    
## college                 -1.791e+01  2.504e+00  -7.154 8.65e-13 ***
## parking_exit             8.215e+00  2.895e+00   2.838 0.004550 ** 
## bar.restaurant           8.645e+00  4.652e+00   1.858 0.063153 .  
## car_rental              -3.702e+00  2.572e+00  -1.439 0.150176    
## coworking               -9.213e+00  4.423e+00  -2.083 0.037264 *  
## shelter                 -1.181e+01  1.137e+00 -10.383  < 2e-16 ***
## bureau_de_change         1.952e+01  3.218e+00   6.067 1.33e-09 ***
## food_cart               -2.892e+01  3.849e+00  -7.514 5.96e-14 ***
## school..historic.       -3.374e-02  2.706e+00  -0.012 0.990052    
## border_control          -5.803e+00  4.617e+00  -1.257 0.208836    
## check_cashing           -4.420e+00  3.717e+00  -1.189 0.234382    
## nail_salon               2.141e+01  5.505e+00   3.889 0.000101 ***
## tax                     -1.264e+00  8.204e+00  -0.154 0.877575    
## catering                 1.268e+01  3.078e+00   4.119 3.82e-05 ***
## bus_station              2.343e+01  2.462e+00   9.518  < 2e-16 ***
## toilets                 -7.616e+00  1.545e+00  -4.929 8.33e-07 ***
## marker                   2.988e+00  2.180e+00   1.371 0.170501    
## social_facility         -1.084e+01  3.709e+00  -2.922 0.003486 ** 
## taxi                    -1.176e+01  2.126e+00  -5.529 3.25e-08 ***
## emergency_phone          1.223e+01  3.361e+00   3.639 0.000274 ***
## fitness_center           6.791e+00  3.383e+00   2.007 0.044725 *  
```

```
## townhall                  -1.294e+01  3.205e+00  -4.038 5.40e-05 ***
## traffic_signals            1.150e-01  1.911e-02   6.020 1.77e-09 ***
## crossing                   1.170e-01  1.516e-02   7.720 1.21e-14 ***
## motorway_junction         -6.659e-01  2.325e-01  -2.864 0.004187 **
## bus_stop                   5.639e-01  1.900e-01   2.969 0.002994 **
## speed_camera               1.061e+01  1.878e+00   5.651 1.62e-08 ***
## stop                      -7.833e-02  1.063e-01  -0.737 0.461114
## turning_circle            -2.395e+00  6.908e-01  -3.467 0.000527 ***
## elevator                  -2.051e-01  4.537e+00  -0.045 0.963940
## traffic_signals.bus_stop -7.458e+00  2.318e+00  -3.217 0.001296 **
## mini_roundabout           -1.218e+01  1.687e+00  -7.217 5.48e-13 ***
## footway                   -8.570e+00  2.168e+00  -3.953 7.74e-05 ***
## street_lamp                7.181e+00  1.800e+00   3.989 6.65e-05 ***
## hotel                      2.536e+00  2.800e-01   9.058  < 2e-16 ***
## artwork                    6.659e-02  1.377e-01   0.484 0.628656
## information               -6.447e+00  8.382e-01  -7.691 1.52e-14 ***
## museum                    -2.020e+00  7.441e-01  -2.714 0.006643 **
## sculpture                 -1.064e+00  6.442e-01  -1.652 0.098473 .
## hostel                    -3.463e+01  4.447e+00  -7.789 7.04e-15 ***
## picnic_site                9.181e-01  8.136e-01   1.128 0.259135
## tour_guide                -6.177e+00  4.228e+00  -1.461 0.144003
## attraction                -3.069e+00  5.701e-01  -5.384 7.36e-08 ***
## landmark                  -1.171e+01  2.045e+00  -5.725 1.05e-08 ***
## motel                     -8.991e+00  2.848e+00  -3.157 0.001597 **
## guest_house               -1.726e+01  3.599e+00  -4.796 1.63e-06 ***
## gallery                   -1.239e-01  1.102e+00  -0.112 0.910451
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.85 on 23262 degrees of freedom
## Multiple R-squared:  0.5309, Adjusted R-squared:  0.5283
## F-statistic: 207.3 on 127 and 23262 DF,  p-value: < 2.2e-16
```

```
# now 'is_work_day1' is NA, what gives?! remember the assumptions of linear
# regression. our covariates must be independent - that is, not correlated. in
# this case if you know the values of weekday, you know the value of
# is_work_day so that assumption doesn't hold. get rid of it!
data_to_model$is_work_day = NULL

# try the model again
model = lm(rentals ~ ., data = data_to_model)
summary(model)
```

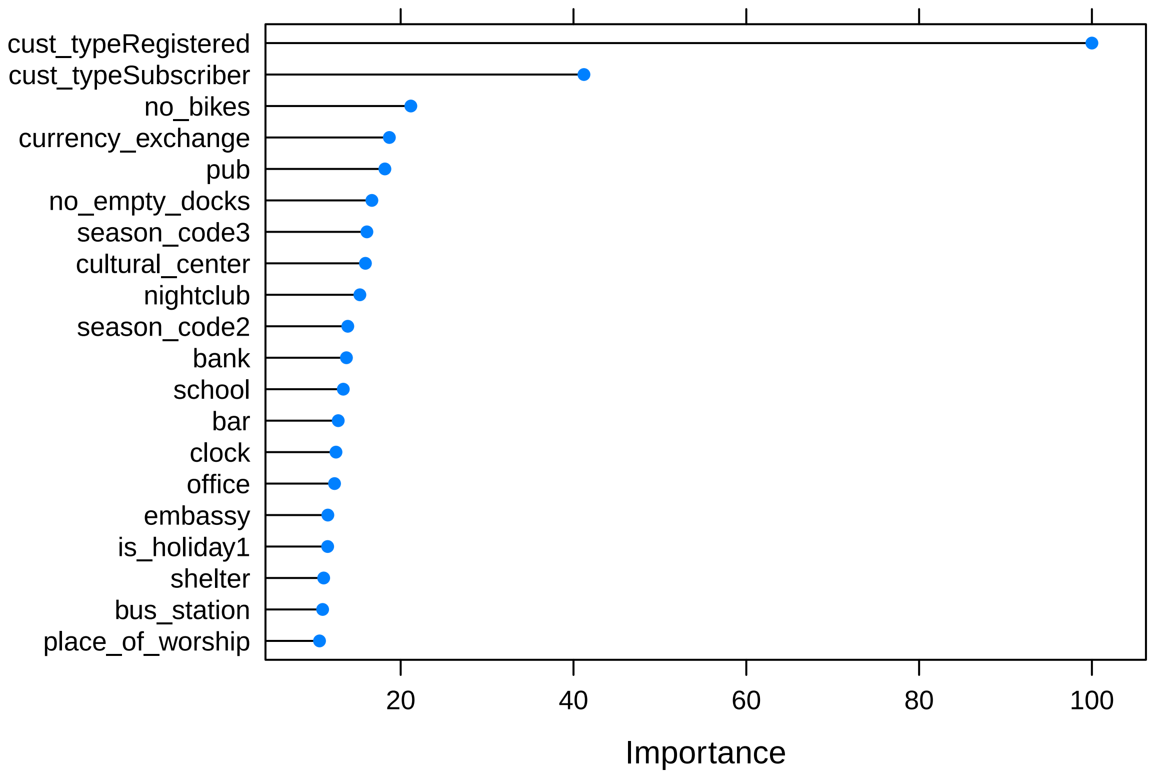
```
##
## Call:
## lm(formula = rentals ~ ., data = data_to_model)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -68.437 -11.258  -0.955   8.856 216.911
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -2.735e+01  1.835e+00 -14.908  < 2e-16 ***
```

```
## cust_typeRegistered       2.732e+01  2.849e-01  95.899  < 2e-16 ***
## cust_typeSubscriber       1.615e+01  4.026e-01  40.109  < 2e-16 ***
## weekday1                  2.304e+00  4.708e-01   4.893 1.00e-06 ***
## weekday2                  1.862e+00  4.875e-01   3.820 0.000134 ***
## weekday3                  2.734e+00  4.976e-01   5.495 3.95e-08 ***
## weekday4                  2.244e+00  4.886e-01   4.593 4.40e-06 ***
## weekday5                  4.683e+00  5.130e-01   9.129  < 2e-16 ***
## weekday6                  3.474e+00  4.690e-01   7.408 1.33e-13 ***
## season_code2             -7.723e+00  6.074e-01 -12.715  < 2e-16 ***
## season_code3             -1.244e+01  8.393e-01 -14.820  < 2e-16 ***
## season_code4             -4.512e+00  4.975e-01  -9.069  < 2e-16 ***
## is_holiday1              -5.004e+00  4.779e-01 -10.472  < 2e-16 ***
## weather_code2            -2.382e+00  4.441e-01  -5.363 8.28e-08 ***
## weather_code3            -9.405e+00  9.631e-01  -9.766  < 2e-16 ***
## duration                  8.655e-04  4.863e-04   1.780 0.075122 .
## temp                     -9.362e+01  1.273e+01  -7.353 2.00e-13 ***
## subjective_temp           1.369e+02  1.437e+01   9.525  < 2e-16 ***
## humidity                 -2.456e+00  2.304e+00  -1.066 0.286593
## windspeed                -1.303e+01  2.845e+00  -4.581 4.66e-06 ***
## no_bikes                  8.103e-01  3.893e-02  20.814  < 2e-16 ***
## no_empty_docks            6.043e-01  3.961e-02  15.257  < 2e-16 ***
## fast_food                 4.193e-01  1.614e-01   2.598 0.009379 **
## parking                  -1.338e+00  3.234e-01  -4.137 3.53e-05 ***
## restaurant                1.306e-01  6.441e-02   2.027 0.042667 *
## convenience              -2.416e+01  2.967e+00  -8.143 4.05e-16 ***
## post_office               2.378e-01  6.769e-01   0.351 0.725334
## bicycle_parking          -2.052e+00  5.059e-01  -4.056 5.02e-05 ***
## drinking_water            1.743e+00  3.797e-01   4.591 4.43e-06 ***
## recycling                -5.961e+00  1.710e+00  -3.487 0.000490 ***
## waste_basket              1.201e+00  2.490e-01   4.825 1.41e-06 ***
## cafe                      7.180e-01  2.546e-01   2.821 0.004798 **
## currency_exchange         4.004e+01  2.468e+00  16.224  < 2e-16 ***
## fountain                  3.241e+00  6.308e-01   5.138 2.80e-07 ***
## pharmacy                  1.102e+00  6.342e-01   1.738 0.082190 .
## car_sharing               2.585e+00  1.077e+00   2.399 0.016426 *
## alcohol                   3.801e+00  2.335e+00   1.627 0.103650
## bank                     -2.106e+00  1.672e-01 -12.600  < 2e-16 ***
## bar                       1.639e+00  1.458e-01  11.247  < 2e-16 ***
## club                     -1.305e+01  3.061e+00  -4.264 2.01e-05 ***
## embassy                   8.440e+00  7.487e-01  11.273  < 2e-16 ***
## food_court               -1.288e+01  1.962e+00  -6.563 5.37e-11 ***
## government                6.668e+00  3.834e+00   1.739 0.082017 .
## internal_kindergarten     2.661e+01  4.383e+00   6.071 1.29e-09 ***
## kindergarten              3.185e-02  4.280e-01   0.074 0.940680
## place_of_worship          1.496e+00  1.397e-01  10.707  < 2e-16 ***
## post_box                 -1.668e+00  5.183e-01  -3.218 0.001293 **
## pub                       4.263e+00  2.759e-01  15.451  < 2e-16 ***
## fuel                      1.376e+00  8.465e-01   1.625 0.104155
## grave_yard                2.618e+00  7.642e-01   3.426 0.000614 ***
## public_building           3.007e+00  6.079e-01   4.947 7.61e-07 ***
## school                    2.194e+00  1.722e-01  12.745  < 2e-16 ***
## fire_station             -4.472e-01  8.551e-01  -0.523 0.601037
## nightclub                -1.675e+01  1.258e+00 -13.310  < 2e-16 ***
## atm                      -4.184e+00  1.262e+00  -3.315 0.000918 ***
```

```
## hospital                 4.134e+00  6.920e-01   5.974 2.34e-09 ***
## doctors                  3.499e+00  6.174e-01   5.667 1.47e-08 ***
## theatre                  3.360e+00  7.419e-01   4.529 5.95e-06 ***
## university               4.302e-01  8.893e-01   0.484 0.628569
## clock                    5.920e+01  5.345e+00  11.075  < 2e-16 ***
## parking_entrance         2.608e-01  1.974e+00   0.132 0.894903
## police                  -5.191e+00  8.388e-01  -6.189 6.16e-10 ***
## cultural_center          7.667e+01  4.995e+00  15.348  < 2e-16 ***
## stripclub               -4.043e-01  1.741e+00  -0.232 0.816356
## marketplace              8.263e+00  1.125e+00   7.347 2.10e-13 ***
## dry_cleaner             -4.864e+00  1.875e+00  -2.594 0.009493 **
## bicycle_repair_station  -5.854e+00  1.045e+00  -5.603 2.13e-08 ***
## office                   3.302e+01  3.003e+00  10.996  < 2e-16 ***
## arts_centre              4.082e+00  1.221e+00   3.343 0.000830 ***
## library                 -2.727e+00  1.105e+00  -2.468 0.013599 *
## studio                  -1.853e+00  2.647e+00  -0.700 0.483849
## strip_club               9.008e+00  2.747e+00   3.279 0.001043 **
## tourist                  8.292e+00  9.126e-01   9.086  < 2e-16 ***
## veterinary               4.297e+00  2.230e+00   1.927 0.053990 .
## community_centre         1.978e+01  2.845e+00   6.953 3.67e-12 ***
## compressed_air           1.316e+01  1.742e+00   7.557 4.28e-14 ***
## tutor                    3.795e+00  1.931e+00   1.966 0.049355 *
## clinic                   1.310e+00  1.186e+00   1.104 0.269568
## dentist                 -5.294e+00  1.474e+00  -3.592 0.000329 ***
## bench                   -1.307e+00  1.611e-01  -8.117 5.00e-16 ***
## cinema                  -1.815e+00  2.678e+00  -0.678 0.497841
## college                 -1.791e+01  2.504e+00  -7.154 8.65e-13 ***
## parking_exit             8.215e+00  2.895e+00   2.838 0.004550 **
## bar.restaurant           8.645e+00  4.652e+00   1.858 0.063153 .
## car_rental              -3.702e+00  2.572e+00  -1.439 0.150176
## coworking               -9.213e+00  4.423e+00  -2.083 0.037264 *
## shelter                 -1.181e+01  1.137e+00 -10.383  < 2e-16 ***
## bureau_de_change         1.952e+01  3.218e+00   6.067 1.33e-09 ***
## food_cart               -2.892e+01  3.849e+00  -7.514 5.96e-14 ***
## school..historic.       -3.374e-02  2.706e+00  -0.012 0.990052
## border_control          -5.803e+00  4.617e+00  -1.257 0.208836
## check_cashing           -4.420e+00  3.717e+00  -1.189 0.234382
## nail_salon               2.141e+01  5.505e+00   3.889 0.000101 ***
## tax                     -1.264e+00  8.204e+00  -0.154 0.877575
## catering                 1.268e+01  3.078e+00   4.119 3.82e-05 ***
## bus_station              2.343e+01  2.462e+00   9.518  < 2e-16 ***
## toilets                 -7.616e+00  1.545e+00  -4.929 8.33e-07 ***
## marker                   2.988e+00  2.180e+00   1.371 0.170501
## social_facility         -1.084e+01  3.709e+00  -2.922 0.003486 **
## taxi                    -1.176e+01  2.126e+00  -5.529 3.25e-08 ***
## emergency_phone          1.223e+01  3.361e+00   3.639 0.000274 ***
## fitness_center           6.791e+00  3.383e+00   2.007 0.044725 *
## townhall                -1.294e+01  3.205e+00  -4.038 5.40e-05 ***
## traffic_signals          1.150e-01  1.911e-02   6.020 1.77e-09 ***
## crossing                 1.170e-01  1.516e-02   7.720 1.21e-14 ***
## motorway_junction       -6.659e-01  2.325e-01  -2.864 0.004187 **
## bus_stop                 5.639e-01  1.900e-01   2.969 0.002994 **
## speed_camera             1.061e+01  1.878e+00   5.651 1.62e-08 ***
## stop                    -7.833e-02  1.063e-01  -0.737 0.461114
```

```
## turning_circle           -2.395e+00  6.908e-01  -3.467 0.000527 ***
## elevator                 -2.051e-01  4.537e+00  -0.045 0.963940
## traffic_signals.bus_stop -7.458e+00  2.318e+00  -3.217 0.001296 **
## mini_roundabout          -1.218e+01  1.687e+00  -7.217 5.48e-13 ***
## footway                  -8.570e+00  2.168e+00  -3.953 7.74e-05 ***
## street_lamp               7.181e+00  1.800e+00   3.989 6.65e-05 ***
## hotel                     2.536e+00  2.800e-01   9.058  < 2e-16 ***
## artwork                   6.659e-02  1.377e-01   0.484 0.628656
## information              -6.447e+00  8.382e-01  -7.691 1.52e-14 ***
## museum                   -2.020e+00  7.441e-01  -2.714 0.006643 **
## sculpture                -1.064e+00  6.442e-01  -1.652 0.098473 .
## hostel                   -3.463e+01  4.447e+00  -7.789 7.04e-15 ***
## picnic_site               9.181e-01  8.136e-01   1.128 0.259135
## tour_guide               -6.177e+00  4.228e+00  -1.461 0.144003
## attraction               -3.069e+00  5.701e-01  -5.384 7.36e-08 ***
## landmark                 -1.171e+01  2.045e+00  -5.725 1.05e-08 ***
## motel                    -8.991e+00  2.848e+00  -3.157 0.001597 **
## guest_house              -1.726e+01  3.599e+00  -4.796 1.63e-06 ***
## gallery                  -1.239e-01  1.102e+00  -0.112 0.910451
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.85 on 23262 degrees of freedom
## Multiple R-squared:  0.5309, Adjusted R-squared:  0.5283
## F-statistic: 207.3 on 127 and 23262 DF,  p-value: < 2.2e-16
```

```r
# ok, we've successfully hit a model but boy does it have a lot of predictors
# lets start evaluating the predictive accuracy

# select the training observations
in_train = createDataPartition(y = data_to_model$rentals,
                               p = 0.75,
                               list = FALSE)

train = data_to_model[in_train, ]
test = data_to_model[-in_train, ]

# when we train with the lm function, we get the same results as using lm()
model_fit = train(rentals ~ .,
                  data = train,
                  method = 'lm',
                  metric = 'RMSE')

# view the relative importance of the predictors
plot(varImp(model_fit), top = 20)
```

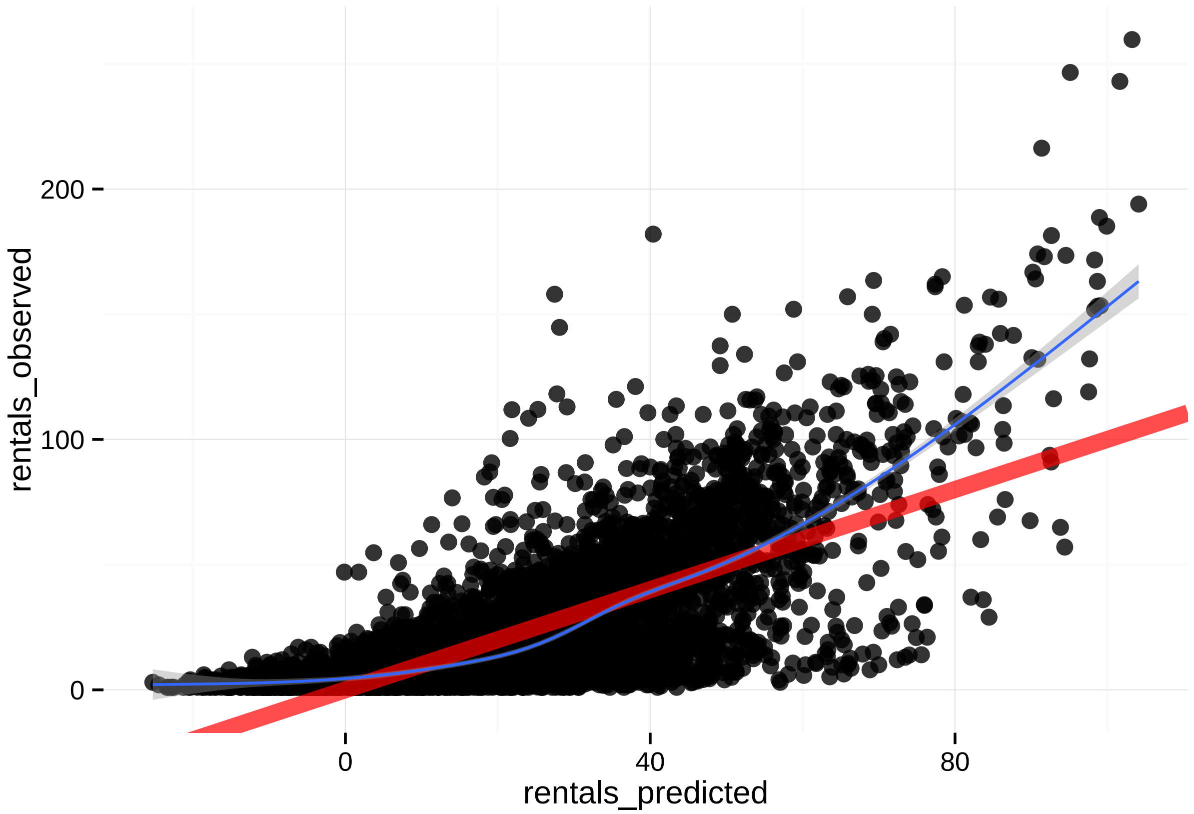



```
rentals_predicted = predict(model_fit, newdata = test)

prediction = data.frame(rentals_predicted,
                        rentals_observed = test$rentals,
                        error = rentals_predicted - test$rentals)
summary(prediction$error)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -156.5000   -9.2340    0.6454   -0.1280   10.8400   61.5900
```

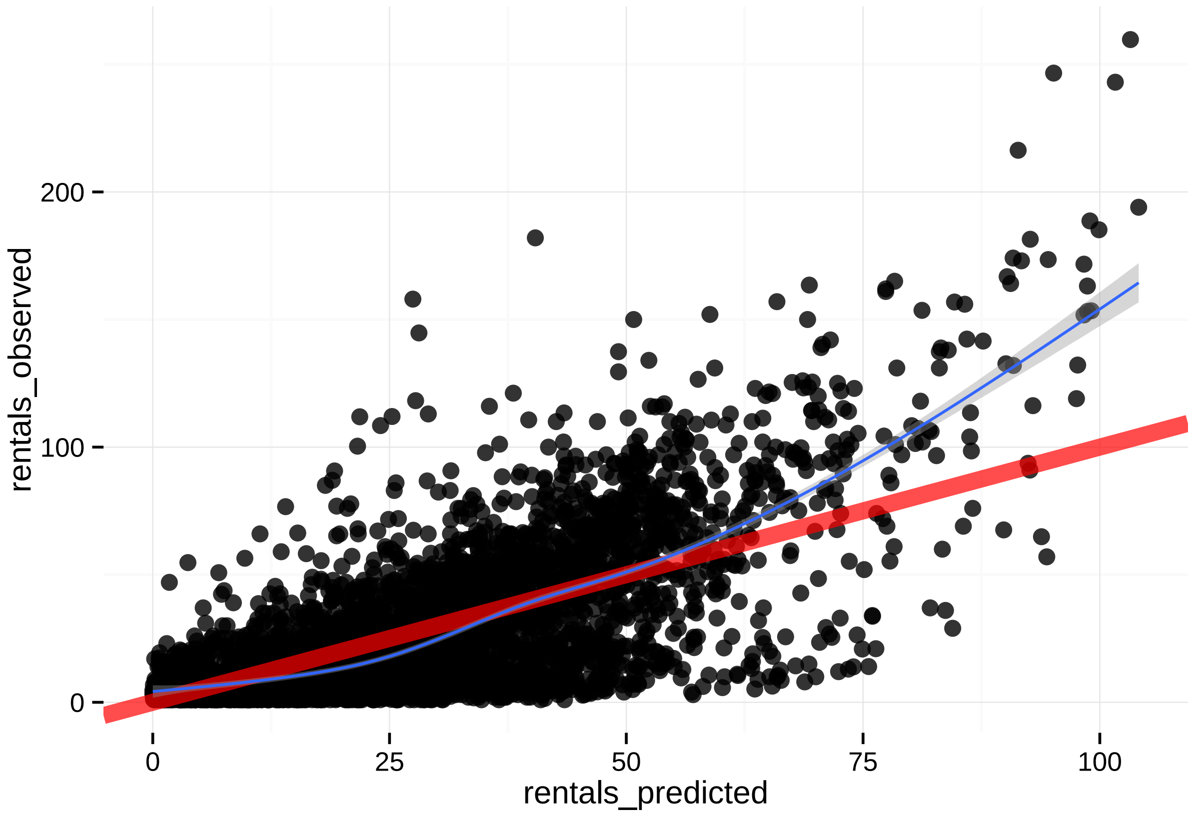
```
ggplot(data = prediction, aes(x = rentals_predicted, y = rentals_observed)) +
  geom_point(size = 3, alpha = 0.80) +
  geom_abline(aes(intercept = 0, slope = 1),
              size = 3, alpha = 0.70, color = 'red') +
  geom_smooth() +
  theme_minimal()
```

```
## geom_smooth: method="auto" and size of largest group is >=1000, so using gam with formula: y ~ s(x, b
```



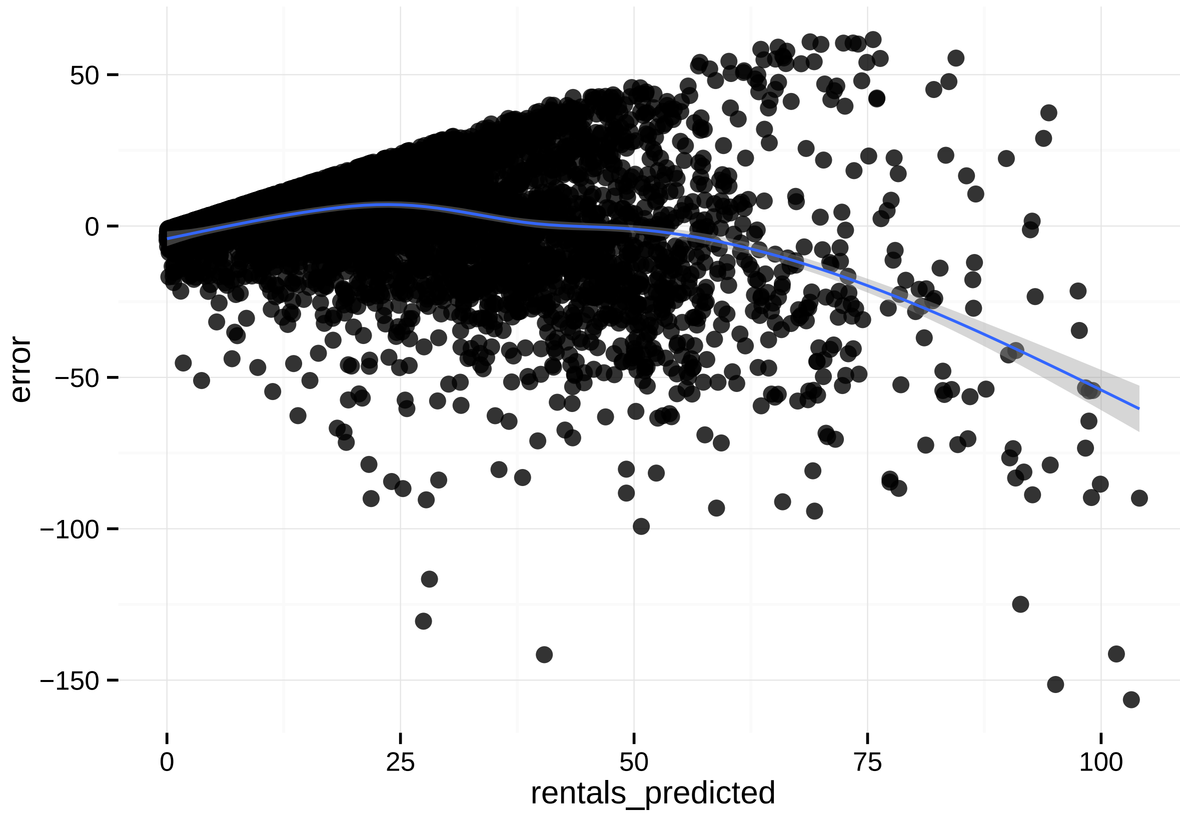

```
ggplot(data = filter(prediction, rentals_predicted > 0), aes(x = rentals_predicted, y = rentals_observe
  geom_point(size = 3, alpha = 0.80) +
  geom_abline(aes(intercept = 0, slope = 1),
              size = 3, alpha = 0.70, color = 'red') +
  geom_smooth() +
  theme_minimal()
```

```
## geom_smooth: method="auto" and size of largest group is >=1000, so using gam with formula: y ~ s(x, 
```




```
ggplot(data = filter(prediction, rentals_predicted > 0), aes(x = rentals_predicted, y = error)) +
  geom_point(size = 3, alpha = 0.80) +
  geom_smooth() +
  theme_minimal()
```

```
## geom_smooth: method="auto" and size of largest group is >=1000, so using gam with formula: y ~ s(x, b
```



50
0
−50
−100
−150
error
0
25
50
75
100
rentals_predicted