Transforming Data

Pedram Navid

September 01, 2016

Transforming Data

Overview

Why transform?

> *Data rarely comes in the format you want it Even if it does, you always want to do something with it Good work upfront saves headache down the road, so don't skimp!*

Tidy Data

Data should be tidy when working with R. We have a tendency to create untidy data, especially in finance.

"tidy datasets are all alike but every messy dataset is messy in its own way"

Tidy data is:

- Each variable is its own column
- Each observation is in its own row

First Steps

Use *dplyr* and *tidyr* to help clean and manipulate data. Check out the cheatsheet: Help > Cheatsheets > Data Manipulation with dplyr,tidyr

Syntax

```r
# Load required packages
library(dplyr)
library(tidyr)

# Let's say we have a dataset
iris
```

```
##     Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1            5.1         3.5          1.4         0.2
## 2            4.9         3.0          1.4         0.2
## 3            4.7         3.2          1.3         0.2
## 4            4.6         3.1          1.5         0.2
## 5            5.0         3.6          1.4         0.2
## 6            5.4         3.9          1.7         0.4
## 7            4.6         3.4          1.4         0.3
## 8            5.0         3.4          1.5         0.2
## 9            4.4         2.9          1.4         0.2
## 10           4.9         3.1          1.5         0.1
```

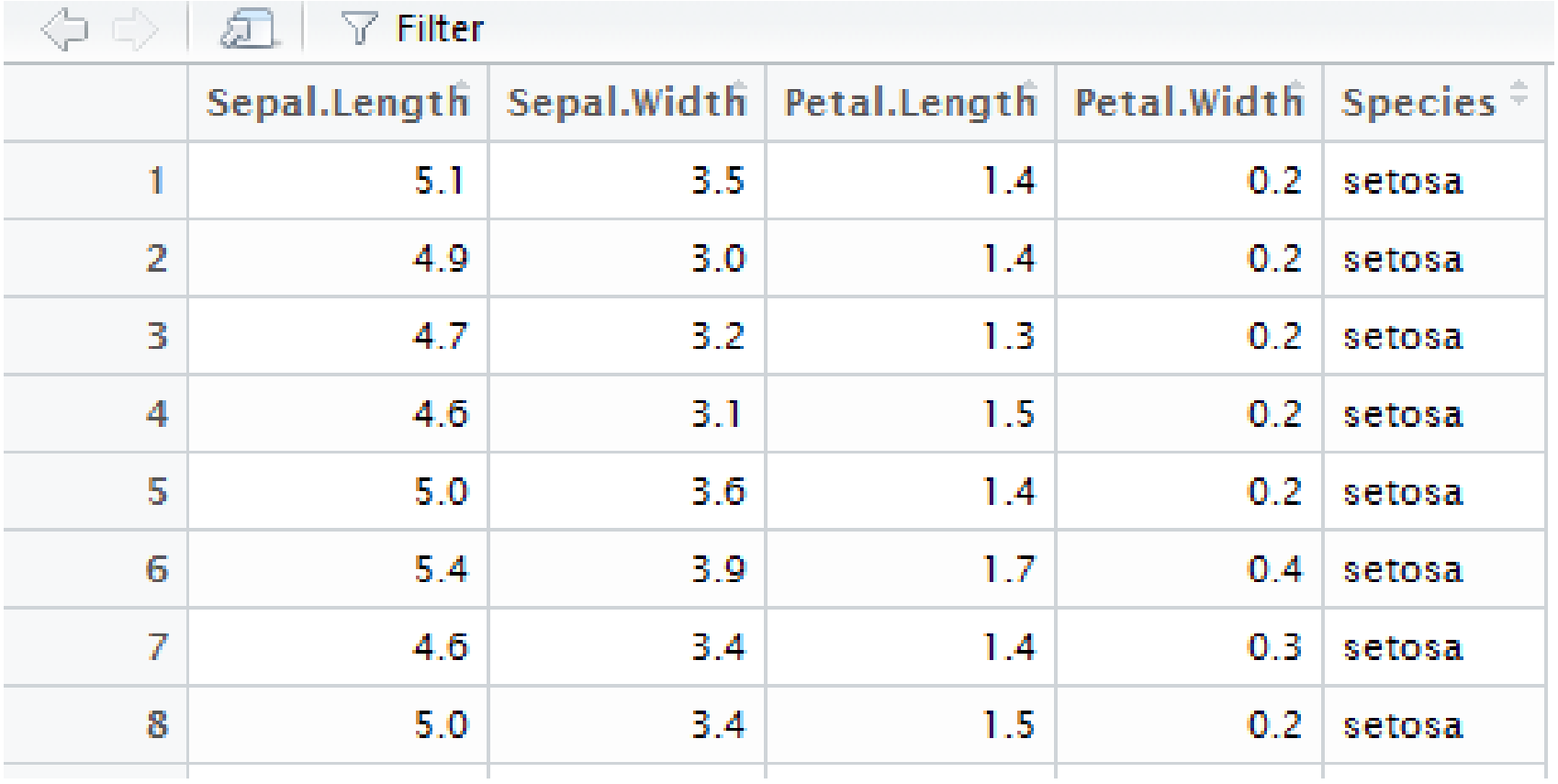
A better way: syntax continued

```
iris_easy <- tbl_df(iris)
iris_easy
```

```
## # A tibble: 150 × 5
##    Sepal.Length Sepal.Width Petal.Length Petal.Width Spe
##           <dbl>       <dbl>        <dbl>       <dbl>  <f
## 1           5.1         3.5          1.4         0.2  se
## 2           4.9         3.0          1.4         0.2  se
## 3           4.7         3.2          1.3         0.2  se
## 4           4.6         3.1          1.5         0.2  se
## 5           5.0         3.6          1.4         0.2  se
## 6           5.4         3.9          1.7         0.4  se
## 7           4.6         3.4          1.4         0.3  se
## 8           5.0         3.4          1.5         0.2  se
## 9           4.4         2.9          1.4         0.2  se
## 10          4.9         3.1          1.5         0.1  se
## # ... with 140 more rows
```

View

```
View(iris_easy)
```



Filter

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa

Figure 1:

Tidy Data

An example:

Is this data tidy? Why/why not?

```
library(readxl)
wide_sales <- read_excel("Transforming_Data/data/wide_data.
head(wide_sales)
```

```
## # A tibble: 4 × 13
##   Region March April   May  June  July August September
##    <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>     <dbl>
## 1  North  5317  2858  4920  5787  2119   2439      3041
## 2   East  4850  1023  6169  8815  6042   4366      3869
## 3   West  4249  6761   100  5685  8408   3427      8939
## 4  South  4195  3285  1654  5257  4206   2144      8361
## # ... with 3 more variables: December <dbl>, January <db
```

Reshaping data: gather

```
tidy_sales <- wide_sales %>%
  gather("month", "sales", 2:13)
head(tidy_sales)
```

```
## # A tibble: 6 × 3
##   Region month sales
##    <chr> <chr> <dbl>
## 1  North March  5317
## 2   East March  4850
## 3   West March  4249
## 4  South March  4195
## 5  North April  2858
## 6   East April  1023
```

Reshaping data: spread

```
tidy_sales %>%
  spread(Region, sales)
```

```
## # A tibble: 12 × 5
##        month  East North South  West
## *      <chr> <dbl> <dbl> <dbl> <dbl>
## 1      April  1023  2858  3285  6761
## 2     August  4366  2439  2144  3427
## 3   December  9792  8880  9652  4974
## 4   February  9746  4526  6991  5139
## 5    January  4806  3644   983   766
## 6       July  6042  2119  4206  8408
## 7       June  8815  5787  5257  5685
## 8      March  4850  5317  4195  4249
## 9        May  6169  4920  1654   100
## 10  November  2635  5128  7307  2893
## 11   October  7819  3811  7400  8959
## 12 September  3869  3041  8361  8939
```

Manipulate Data

Intro to Dplyr

You will spend more time with dplyr than any other R package. First thing to understand is the pipe operator: %>%

%>% passes the result of the previous line to the next line, making your code much more readable

Consider this fake example:

```
miss_muffet_sitting <- sit(miss_muffet, tuffet)
miss_muffet_eating <- eat(miss_muffet_eating, curds_and_whe
muffet_spider <- sit(miss_muffet_eating, spider, where = 'b
muffet_away <- run(muffet_spider, reason = 'frightened')

# Or..

miss_muffet %>%
  sit(tuffet) %>%
  eat(curds_and_whey) %>%
  sit(spider, where = 'beside') %>%
  run(reason = 'frightened')
```

dplyr: continued

Best way to learn it is to use it. . . .a lot.

```
library(dplyr)
parking <- read.csv("Transforming_Data/data/parking_data.cs
  tbl_df()
parking
```

```
## # A tibble: 234,983 × 11
##    tag_number_masked date_of_infraction infraction_code
##                <fctr>              <int>           <int>
## 1            ***42781           20141123             210
## 2            ***30955           20141123             210
## 3            ***57421           20141123             210
## 4            ***14417           20141123             210
## 5            ***35411           20141123             207
## 6            ***49340           20141123               5
## 7            ***17984           20141123               3
## 8            ***30956           20141123             207
```

dplyr: select

Pick columns, and optionally rename and reorder them

```
# Pick
parking %>%
  select(date_of_infraction, infraction_description, 5) %>%
  head()
```

```
## # A tibble: 6 × 3
##   date_of_infraction                infraction_description set_
##                <int>                                 <fctr>
## 1           20141123   PARK FAIL TO DISPLAY RECEIPT
## 2           20141123   PARK FAIL TO DISPLAY RECEIPT
## 3           20141123   PARK FAIL TO DISPLAY RECEIPT
## 4           20141123   PARK FAIL TO DISPLAY RECEIPT
## 5           20141123  PARK FAIL TO DEP. FEE MACHINE
## 6           20141123 PARK-SIGNED HWY-PROHIBIT DY/TM
```

```r
# Rename
parking %>%
  select(inf_date = date_of_infraction, desc = infraction_c
  head()
```

```
## # A tibble: 6 × 4
##   inf_date                              desc  fine  time
##      <int>                             <fctr> <int> <int>
## 1 20141123   PARK FAIL TO DISPLAY RECEIPT    30  1554
## 2 20141123   PARK FAIL TO DISPLAY RECEIPT    30  1554
## 3 20141123   PARK FAIL TO DISPLAY RECEIPT    30  1554
## 4 20141123   PARK FAIL TO DISPLAY RECEIPT    30  1554
## 5 20141123  PARK FAIL TO DEP. FEE MACHINE    30  1554
## 6 20141123 PARK-SIGNED HWY-PROHIBIT DY/TM    40  1554
```

```r
# More picking
parking %>%
  select(starts_with('location')) %>%
  head()
```

```
## # A tibble: 6 × 4
##   location1           location2 location3    location4
##       <fctr>              <fctr>    <fctr>        <fctr>
## 1       N/S         LOURDES LANE       W/O HOMEWOOD AVE
## 2        NR       391 KING ST W
## 3        NR    96 ST PATRICK ST
## 4       OPP      986 BLOOR ST W
## 5        NR    111 ELIZABETH ST
## 6        NR 56 BISHOP TUTU BLVD
```

```
parking %>%
  select(-starts_with('location')) %>%
  head()
```

```
## # A tibble: 6 × 7
##   tag_number_masked date_of_infraction infraction_code
##               <fctr>              <int>            <int>
## 1          ***42781           20141123              210
## 2          ***30955           20141123              210
## 3          ***57421           20141123              210
## 4          ***14417           20141123              210
## 5          ***35411           20141123              207
## 6          ***49340           20141123                5
## # ... with 4 more variables: infraction_description <fct
## #   set_fine_amount <int>, time_of_infraction <int>, pro
```

Dplyr: Filter

Subset data using logic

```
parking %>%
  filter(set_fine_amount > 300)
```

```
## # A tibble: 1,270 × 11
##    tag_number_masked date_of_infraction infraction_code
##                <fctr>              <int>           <int>
## 1           ***54560           20141123             355
## 2           ***49346           20141123             367
## 3           ***61327           20141123             355
## 4           ***30223           20141123             355
## 5           ***13649           20141123             367
## 6           ***55315           20141123             363
## 7           ***49375           20141123             367
## 8           ***88513           20141123             355
## 9           ***03453           20141123             355
## 10          ***53656           20141124             363
```

```
parking %>%
  filter(infraction_code == 355, province != "ON")
```

```
## # A tibble: 18 × 11
##    tag_number_masked date_of_infraction infraction_code
##                <fctr>              <int>           <int>
## 1            ***45827           20141127             355
## 2            ***68860           20141127             355
## 3            ***69264           20141128             355
## 4            ***11902           20141129             355
## 5            ***79981           20141202             355
## 6            ***11996           20141204             355
## 7            ***42247           20141214             355
## 8            ***45845           20141214             355
## 9            ***80547           20141215             355
## 10           ***44391           20141216             355
## 11           ***84360           20141216             355
## 12           ***69611           20141218             355
## 13           ***48790           20141218             355
```

Distinct

```
parking %>%
  select(infraction_code, infraction_description) %>%
  distinct()
```

```
## # A tibble: 146 × 2
##    infraction_code         infraction_description
##              <int>                         <fctr>
## 1              210   PARK FAIL TO DISPLAY RECEIPT
## 2              207  PARK FAIL TO DEP. FEE MACHINE
## 3                5 PARK-SIGNED HWY-PROHIBIT DY/TM
## 4                3       PARK ON PRIVATE PROPERTY
## 5                8   STAND VEH.-PROHIBIT TIME/DAY
## 6               16  PARK-WITHIN 9M INTERSECT ROAD
## 7              347            PARK IN A FIRE ROUTE
## 8              406  PARK-VEH. W/O VALID ONT PLATE
## 9              355   PARK IN ACCESSIBLE NO PERMIT
## 10               6 PARK-SIGNED HWY-EXC PERMT TIME
## # ... with 136 more rows
```

Samples

```
set.seed(42)
parking %>%
  select(location2, infraction_description, set_fine_amount
  sample_n(10, replace = FALSE)
```

```
## # A tibble: 10 × 3
##                location2                infraction_description se
##                   <fctr>                                <fctr>
## 1        ST CLAIR AVE W    PARK-SIGNED HWY-PUBLIC LANE
## 2         225 KING ST W  STOP-SIGNED HIGHWAY-RUSH HOUR
## 3          16 NEWTON DR PARK-SIGNED HWY-PROHIBIT DY/TM
## 4  700 HUMBERWOOD BLVD       PARK ON PRIVATE PROPERTY
## 5        225 QUEBEC AVE PARK PROHIBITED TIME NO PERMIT
## 6        BRENTWOOD RD N   PARK FAIL TO DISPLAY RECEIPT
## 7     20 EGLINTON AVE W   STAND VEH.-PROHIBIT TIME/DAY
## 8        ST NICHOLAS ST PARK-SIGNED HWY-PROHIBIT DY/TM
## 9         14 COLLEGE ST  PARK FAIL TO DEP. FEE MACHINE
## 10        256 DONLEA DR PARK-SIGNED HWY-PROHIBIT DY/TM
```

Slice

```
parking %>%
  slice(12:15)
```

```
## # A tibble: 4 × 11
##   tag_number_masked date_of_infraction infraction_code
##               <fctr>              <int>           <int>
## 1          ***49341           20141123               5
## 2          ***42782           20141123             207
## 3          ***14419           20141123             210
## 4          ***14957           20141123               5
## # ... with 8 more variables: infraction_description <fct
## #   set_fine_amount <int>, time_of_infraction <int>, loc
## #   location2 <fctr>, location3 <fctr>, location4 <fctr>
```

Top

```
head(parking, n = 8)
```

```
## # A tibble: 8 × 11
##   tag_number_masked date_of_infraction infraction_code
##               <fctr>              <int>           <int>
## 1          ***42781           20141123             210
## 2          ***30955           20141123             210
## 3          ***57421           20141123             210
## 4          ***14417           20141123             210
## 5          ***35411           20141123             207
## 6          ***49340           20141123               5
## 7          ***17984           20141123               3
## 8          ***30956           20141123             207
## # ... with 8 more variables: infraction_description <fct
## #   set_fine_amount <int>, time_of_infraction <int>, loc
## #   location2 <fctr>, location3 <fctr>, location4 <fctr>
```

Bottom

```
tail(parking)
```

```
## # A tibble: 6 × 11
##   tag_number_masked date_of_infraction infraction_code
##               <fctr>              <int>           <int>
## 1          ***48217           20141231               3
## 2          ***67949           20141231              29
## 3          ***60555           20141231               5
## 4          ***45157           20141231              29
## 5          ***48218           20141231               3
## 6          ***87324           20141231               9
## # ... with 8 more variables: infraction_description <fct
## #   set_fine_amount <int>, time_of_infraction <int>, loc
## #   location2 <fctr>, location3 <fctr>, location4 <fctr>
```

Top X

```
parking %>%
  filter(province != "ON") %>%
  top_n(n = 10, wt = set_fine_amount) %>%
  select(infraction_description, location2, set_fine_amount
```

```
## # A tibble: 55 × 3
##                infraction_description           location2 se
##                                <fctr>              <fctr>
## 1  STND ONSTRT ACCESSIBLE NO PRMT      150 SUDBURY ST
## 2  PARK ONSTRT ACCESSIBLE NO PRMT   143 GORE VALE AVE
## 3  STND ONSTRT ACCESSIBLE NO PRMT         101 ELM ST
## 4  STND ONSTRT ACCESSIBLE NO PRMT      410 COLLEGE ST
## 5  STND ONSTRT ACCESSIBLE NO PRMT      410 COLLEGE ST
## 6    PARK IN ACCESSIBLE NO PERMIT    3401 DUFFERIN ST
## 7    PARK IN ACCESSIBLE NO PERMIT 2020 SHEPPARD AVE W
## 8    PARK IN ACCESSIBLE NO PERMIT   90 EGLINTON AVE E
## 9  STND ONSTRT ACCESSIBLE NO PRMT      410 COLLEGE ST
## 10 STND ONSTRT ACCESSIBLE NOT D/O         8 SULTAN ST
```

Mutate: make new variables

```
library(stringr)
parking %>%
  select(2:8, -infraction_code) %>%
  mutate(lower_location = str_to_lower(location2))
```

```
## # A tibble: 234,983 × 7
##    date_of_infraction         infraction_description set
##                 <int>                          <fctr>
## 1            20141123   PARK FAIL TO DISPLAY RECEIPT
## 2            20141123   PARK FAIL TO DISPLAY RECEIPT
## 3            20141123   PARK FAIL TO DISPLAY RECEIPT
## 4            20141123   PARK FAIL TO DISPLAY RECEIPT
## 5            20141123  PARK FAIL TO DEP. FEE MACHINE
## 6            20141123 PARK-SIGNED HWY-PROHIBIT DY/TM
## 7            20141123       PARK ON PRIVATE PROPERTY
## 8            20141123  PARK FAIL TO DEP. FEE MACHINE
## 9            20141123 PARK-SIGNED HWY-PROHIBIT DY/TM
## 10           20141123   PARK FAIL TO DISPLAY RECEIPT
## # ... with 234,973 more rows, and 4 more variables:
```

rank things

```
parking %>%
  select(tag_number_masked, set_fine_amount) %>%
  mutate(rank = min_rank(-set_fine_amount)) %>%
  arrange(rank)
```

```
## # A tibble: 234,983 × 3
##    tag_number_masked set_fine_amount  rank
##                <fctr>           <int> <int>
## 1           ***54560             450     1
## 2           ***49346             450     1
## 3           ***61327             450     1
## 4           ***30223             450     1
## 5           ***13649             450     1
## 6           ***55315             450     1
## 7           ***49375             450     1
## 8           ***88513             450     1
## 9           ***03453             450     1
## 10          ***53656             450     1
```

Group By / Summary

```
parking %>%
  summarise(total_fee = sum(set_fine_amount))
```

```
## # A tibble: 1 × 1
##   total_fee
##       <int>
## 1  10885905
```

```
parking %>%
  group_by(province) %>%
  summarise(total_fee = sum(set_fine_amount)) %>%
  arrange(desc(total_fee)) %>%
  mutate(rank = min_rank(-total_fee))
```

```
## # A tibble: 65 × 3
##    province total_fee  rank
##      <fctr>     <int> <int>
## 1        ON  10300590     1
```

more summary:

```
parking %>%
  group_by(province) %>%
  summarise(mean_amt = mean(set_fine_amount),
    max_amt = max(set_fine_amount),
    n = n(),
    sd = sd(set_fine_amount)) %>%
  arrange(desc(mean_amt))
```

```
## # A tibble: 65 × 5
##    province mean_amt max_amt     n       sd
##      <fctr>    <dbl>   <int> <int>    <dbl>
## 1        TX 63.80368     450   163 78.17129
## 2        VT 61.81818     150    11 47.92039
## 3        OR 57.17391     450    46 68.74795
## 4        NC 56.15385     450    39 72.60620
## 5        AB 55.82815     450  1769 55.38691
## 6        IN 55.07812     450    64 62.43404
## 7        LA 54.00000     150    10 40.33196
```

joins

inner joins

```
prov_names <- data.frame(province = c('AB', 'BC', 'MB', 'NB
  'NT', 'NU', 'ON', 'PE', 'QC', 'SK', 'YT'),
  long_name = c('Alberta', 'British Columbia', 'Manitoba',
  'Nova Scotia', 'Northwest Territories', 'Nunavut', 'Ontar
    'Quebec', 'Saskatchewan', 'Yukon'))

small_parking <- parking %>%
  select(province, location2, set_fine_amount)

inner_join(small_parking, prov_names) %>%
  filter(province %in% c('NS', 'NY'))
```

```
## Warning in inner_join_impl(x, y, by$x, by$y, suffix$x, s
## factors with different levels, coercing to character vec
```

```
## # A tibble: 552 × 4
##    province        location2 set_fine_amount   long_name
##       <chr>            <fctr>           <int>      <fctr>
```

left/right joins

```
left_join(small_parking, prov_names) %>%
  filter(province %in% c('NS', 'NY'))
```

```
## Warning in left_join_impl(x, y, by$x, by$y, suffix$x, su
## factors with different levels, coercing to character vec
```

```
## # A tibble: 1,381 × 4
##    province        location2 set_fine_amount   long_name
##       <chr>            <fctr>           <int>      <fctr>
## 1        NY   444 FRONT ST W              30          NA
## 2        NS    82 LINDYLOU RD             40 Nova Scotia
## 3        NY  399 SPADINA AVE              30          NA
## 4        NS       BOGERT AVE              30 Nova Scotia
## 5        NS    125 GEORGE ST              30 Nova Scotia
## 6        NY 92 YORKVILLE AVE              30          NA
## 7        NY    269 MUTUAL ST              30          NA
## 8        NY 228 DANFORTH AVE              30          NA
## 9        NS 26 BROOKFIELD ST              40 Nova Scotia
```

anti joins

```
anti_join(small_parking, prov_names)
```

```
## # A tibble: 4,815 × 3
##    province          location2 set_fine_amount
##       <fctr>              <fctr>           <int>
## 1        AR 1090 DON MILLS RD              250
## 2        AR   2305 QUEEN ST E               30
## 3        AR    19 CHESTER AVE               40
## 4        AR         HARBORD ST              30
## 5        AR   1963 QUEEN ST E               30
## 6        AR   1963 QUEEN ST E               30
## 7        AR  420 DANFORTH AVE               30
## 8        AR     84 SAULTER ST               40
## 9        AR   79 TORBARRIE RD               40
## 10       AR   1947 QUEEN ST E               30
## # ... with 4,805 more rows
```