

Manipulating data

Stewart Li

9/14/2020

Pipe

```
df <- mtcars %>%  
  as_tibble(rownames = "car") %>%  
  mutate(across(c(am, gear, carb), as.factor))
```

```
df %T>%  
  glimpse() %>%  
  arrange(desc(mpg))
```

```
## Rows: 32  
## Columns: 12  
## $ car   <chr> "Mazda RX4", "Mazda RX4 Wag", "Datsun 710", "Hornet 4 Drive", ...  
## $ mpg   <dbl> 21.0, 21.0, 22.8, 21.4, 18.7, 18.1, 14.3, 24.4, 22.8, 19.2, 17...  
## $ cyl   <dbl> 6, 6, 4, 6, 8, 6, 8, 4, 4, 6, 6, 8, 8, 8, 8, 8, 4, 4, 4, 4,...  
## $ disp  <dbl> 160.0, 160.0, 108.0, 258.0, 360.0, 225.0, 360.0, 146.7, 140.8,...  
## $ hp    <dbl> 110, 110, 93, 110, 175, 105, 245, 62, 95, 123, 123, 180, 180, ...  
## $ drat  <dbl> 3.90, 3.90, 3.85, 3.08, 3.15, 2.76, 3.21, 3.69, 3.92, 3.92, 3....  
## $ wt    <dbl> 2.620, 2.875, 2.320, 3.215, 3.440, 3.460, 3.570, 3.190, 3.150,...  
## $ qsec  <dbl> 16.46, 17.02, 18.61, 19.44, 17.02, 20.22, 15.84, 20.00, 22.90,...  
## $ vs    <dbl> 0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1,...  
## $ am    <fct> 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0,...  
## $ gear  <fct> 4, 4, 4, 3, 3, 3, 3, 4, 4, 4, 4, 3, 3, 3, 3, 3, 3, 4, 4, 4, 3,...  
## $ carb  <fct> 4, 4, 1, 1, 2, 1, 4, 2, 2, 4, 4, 3, 3, 3, 4, 4, 4, 1, 2, 1, 1,...
```

```
## # A tibble: 32 x 12  
##   car           mpg   cyl  disp    hp  drat    wt  qsec    vs  am  gear  carb  
##   <chr>         <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <fct> <fct> <fct>  
## 1 Toyota Cor~  33.9     4  71.1   65  4.22  1.84  19.9     1  1     4     1  
## 2 Fiat 128     32.4     4  78.7   66  4.08  2.2   19.5     1  1     4     1  
## 3 Honda Civic  30.4     4  75.7   52  4.93  1.62  18.5     1  1     4     2  
## 4 Lotus Euro~  30.4     4  95.1  113  3.77  1.51  16.9     1  1     5     2  
## 5 Fiat Xl-9    27.3     4   79    66  4.08  1.94  18.9     1  1     4     1  
## 6 Porsche 91~  26      4 120.    91  4.43  2.14  16.7     0  1     5     2  
## 7 Merc 240D    24.4     4 147.    62  3.69  3.19   20      1  0     4     2  
## 8 Datsun 710   22.8     4 108     93  3.85  2.32  18.6     1  1     4     1  
## 9 Merc 230     22.8     4 141.    95  3.92  3.15  22.9     1  0     4     2  
## 10 Toyota Cor~ 21.5     4 120.    97  3.7   2.46  20.0     1  0     3     1  
## # ... with 22 more rows
```

```
df %$%  
  c(mpg, hp)
```

```
## [1] 21.0 21.0 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 17.8 16.4
## [13] 17.3 15.2 10.4 10.4 14.7 32.4 30.4 33.9 21.5 15.5 15.2 13.3
## [25] 19.2 27.3 26.0 30.4 15.8 19.7 15.0 21.4 110.0 110.0 93.0 110.0
## [37] 175.0 105.0 245.0 62.0 95.0 123.0 123.0 180.0 180.0 180.0 205.0 215.0
## [49] 230.0 66.0 52.0 65.0 97.0 150.0 150.0 245.0 175.0 66.0 91.0 113.0
## [61] 264.0 175.0 335.0 109.0
```

Tidy

1. row: each observation.
2. column: variables.
3. cell: value.

Seprate

```
df %>%
  separate(car, into = c('var1', 'var2'), sep = " ", extra = "merge", remove = FALSE, fill = "right")
```

```
## # A tibble: 32 x 14
##   car   var1 var2   mpg   cyl  disp    hp  drat    wt   qsec    vs  am  gear
##   <chr> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <fct> <fct>
## 1 Mazd~ Mazda RX4    21     6  160   110  3.9   2.62  16.5     0  1    4
## 2 Mazd~ Mazda RX4 ~    21     6  160   110  3.9   2.88  17.0     0  1    4
## 3 Dats~ Dats~  710   22.8    4  108    93  3.85   2.32  18.6     1  1    4
## 4 Horn~ Horn~  4 Dr~  21.4    6  258   110  3.08   3.22  19.4     1  0    3
## 5 Horn~ Horn~ Spor~  18.7    8  360   175  3.15   3.44  17.0     0  0    3
## 6 Vali~ Vali~ <NA>   18.1    6  225   105  2.76   3.46  20.2     1  0    3
## 7 Dust~ Dust~  360   14.3    8  360   245  3.21   3.57  15.8     0  0    3
## 8 Merc~ Merc  240D   24.4    4  147.    62  3.69   3.19   20      1  0    4
## 9 Merc~ Merc  230   22.8    4  141.    95  3.92   3.15  22.9     1  0    4
## 10 Merc~ Merc  280   19.2    6  168.   123  3.92   3.44  18.3     1  0    4
## # ... with 22 more rows, and 1 more variable: carb <fct>
```

```
df %>%
  separate_rows(car)
```

```
## # A tibble: 68 x 12
##   car      mpg   cyl  disp    hp  drat    wt   qsec    vs  am   gear  carb
##   <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <fct> <fct> <fct>
## 1 Mazda    21     6   160   110   3.9   2.62  16.5    0  1     4     4
## 2 RX4      21     6   160   110   3.9   2.62  16.5    0  1     4     4
## 3 Mazda    21     6   160   110   3.9   2.88  17.0    0  1     4     4
## 4 RX4      21     6   160   110   3.9   2.88  17.0    0  1     4     4
## 5 Wag      21     6   160   110   3.9   2.88  17.0    0  1     4     4
## 6 Datsun   22.8    4   108    93   3.85   2.32  18.6    1  1     4     1
## 7 710      22.8    4   108    93   3.85   2.32  18.6    1  1     4     1
## 8 Hornet   21.4    6   258   110   3.08   3.22  19.4    1  0     3     1
## 9 4        21.4    6   258   110   3.08   3.22  19.4    1  0     3     1
## 10 Drive   21.4    6   258   110   3.08   3.22  19.4    1  0     3     1
## # ... with 58 more rows
```

Extract

```
df %>%
  tidyr::extract(., car, into = c("var1"), regex = "(\\d\\d?\\d?[-]?\\d?)", remove = FALSE) %>%
  select(car, var1)
```

```
## # A tibble: 32 x 2
##   car          var1
##   <chr>         <chr>
## 1 Mazda RX4      4
## 2 Mazda RX4 Wag  4
## 3 Datsun 710     710
## 4 Hornet 4 Drive 4
## 5 Hornet Sportabout <NA>
## 6 Valiant        <NA>
## 7 Duster 360     360
## 8 Merc 240D      240
## 9 Merc 230       230
## 10 Merc 280      280
## # ... with 22 more rows
```

Gather/Spread

```
df %>%
  pivot_longer(c(mpg, disp, hp),
    names_to = "variables",
    values_to = "values") %>%
  pivot_wider(names_from = variables,
    values_from = values)
```

```
## # A tibble: 32 x 12
##   car          cyl  drat   wt  qsec    vs am  gear  carb   mpg  disp   hp
##   <chr>        <dbl> <dbl> <dbl> <dbl> <dbl> <fct> <fct> <fct> <dbl> <dbl> <dbl>
## 1 Mazda RX4      6  3.9   2.62  16.5   0  1    4     4    21   160   110
## 2 Mazda RX4 ~    6  3.9   2.88  17.0   0  1    4     4    21   160   110
## 3 Datsun 710     4  3.85  2.32  18.6   1  1    4     1   22.8  108    93
## 4 Hornet 4 D~    6  3.08  3.22  19.4   1  0    3     1   21.4  258   110
## 5 Hornet Spo~    8  3.15  3.44  17.0   0  0    3     2   18.7  360   175
## 6 Valiant        6  2.76  3.46  20.2   1  0    3     1   18.1  225   105
## 7 Duster 360     8  3.21  3.57  15.8   0  0    3     4   14.3  360   245
## 8 Merc 240D      4  3.69  3.19   20    1  0    4     2   24.4  147.    62
## 9 Merc 230       4  3.92  3.15  22.9   1  0    4     2   22.8  141.    95
## 10 Merc 280      6  3.92  3.44  18.3   1  0    4     4   19.2  168.   123
## # ... with 22 more rows
```

Rows

Filter

```
df %>%
  filter(am == 0, hp >= 150, hp <= 300)
```

```
## # A tibble: 12 x 12
##   car          mpg  cyl  disp   hp  drat   wt  qsec    vs am  gear  carb
##   <chr>        <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <fct> <fct> <fct>
## 1 Hornet Spo~  18.7     8  360   175  3.15  3.44  17.0   0  0    3     2
## 2 Duster 360   14.3     8  360   245  3.21  3.57  15.8   0  0    3     4
## 3 Merc 450SE   16.4     8  276.   180  3.07  4.07  17.4   0  0    3     3
## 4 Merc 450SL   17.3     8  276.   180  3.07  3.73  17.6   0  0    3     3
## 5 Merc 450SLC  15.2     8  276.   180  3.07  3.78  18     0  0    3     3
## 6 Cadillac F~  10.4     8  472   205  2.93  5.25  18.0   0  0    3     4
## 7 Lincoln Co~  10.4     8  460   215   3     5.42  17.8   0  0    3     4
## 8 Chrysler I~  14.7     8  440   230  3.23  5.34  17.4   0  0    3     4
## 9 Dodge Chal~  15.5     8  318   150  2.76  3.52  16.9   0  0    3     2
## 10 AMC Javelin  15.2     8  304   150  3.15  3.44  17.3   0  0    3     2
## 11 Camaro Z28   13.3     8  350   245  3.73  3.84  15.4   0  0    3     4
## 12 Pontiac Fi~  19.2     8  400   175  3.08  3.84  17.0   0  0    3     2
```

```
df %>%
  group_by(hp > 200) %>%
  filter(mpg == max(mpg))
```

```
## # A tibble: 2 x 13
## # Groups:   hp > 200 [2]
##   car      mpg    cyl  disp    hp  drat    wt  qsec    vs  am    gear  carb
##   <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <fct> <fct> <fct>
## 1 Toyo~  33.9     4  71.1    65  4.22  1.84  19.9     1  1     4     1
## 2 Ford~  15.8     8 351     264  4.22  3.17  14.5     0  1     5     4
## # ... with 1 more variable: `hp > 200` <lgl>
```

```
df %>%
  filter(substr(car, 2, 2) == "a")
```

```
## # A tibble: 7 x 12
##   car      mpg    cyl  disp    hp  drat    wt  qsec    vs  am    gear  carb
##   <chr>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <fct> <fct> <fct>
## 1 Mazda RX4      21     6  160   110  3.9   2.62  16.5     0  1     4     4
## 2 Mazda RX4 W~   21     6  160   110  3.9   2.88  17.0     0  1     4     4
## 3 Datsun 710    22.8     4  108    93  3.85  2.32  18.6     1  1     4     1
## 4 Valiant      18.1     6  225   105  2.76  3.46  20.2     1  0     3     1
## 5 Cadillac Fl~  10.4     8  472   205  2.93  5.25  18.0     0  0     3     4
## 6 Camaro Z28    13.3     8  350   245  3.73  3.84  15.4     0  0     3     4
## 7 Maserati Bo~   15     8  301   335  3.54  3.57  14.6     0  1     5     8
```

Distinct

```
df %>%
  distinct(am, gear, carb) %>% count(am, gear, carb)
```

```
## # A tibble: 13 x 4
##   am    gear  carb     n
##   <fct> <fct> <fct> <int>
## 1 0      3     1     1
## 2 0      3     2     1
## 3 0      3     3     1
## 4 0      3     4     1
## 5 0      4     2     1
## 6 0      4     4     1
## 7 1      4     1     1
## 8 1      4     2     1
## 9 1      4     4     1
## 10 1     5     2     1
## 11 1     5     4     1
## 12 1     5     6     1
## 13 1     5     8     1
```

```
df %>%
  complete(am, gear, carb) %>% count(am, gear, carb)
```

```
## # A tibble: 36 x 4
##   am     gear carb     n
##   <fct> <fct> <fct> <int>
## 1 0       3     1       3
## 2 0       3     2       4
## 3 0       3     3       3
## 4 0       3     4       5
## 5 0       3     6       1
## 6 0       3     8       1
## 7 0       4     1       1
## 8 0       4     2       2
## 9 0       4     3       1
## 10 0      4     4       2
## # ... with 26 more rows
```

Slice

```
df %>%
  arrange(desc(mpg)) %>%
  slice(1:3, (n()-2):n()) # slice_tail(n = 3)
```

```
## # A tibble: 6 x 12
##   car      mpg   cyl  disp    hp  drat    wt   qsec    vs  am   gear  carb
##   <chr>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <fct> <fct> <fct>
## 1 Toyota Coro~ 33.9     4  71.1    65  4.22  1.84  19.9     1  1     4     1
## 2 Fiat 128      32.4     4  78.7    66  4.08  2.2   19.5     1  1     4     1
## 3 Honda Civic   30.4     4  75.7    52  4.93  1.62  18.5     1  1     4     2
## 4 Camaro Z28    13.3     8 350      245  3.73  3.84  15.4     0  0     3     4
## 5 Cadillac Fl~ 10.4     8 472      205  2.93  5.25  18.0     0  0     3     4
## 6 Lincoln Con~ 10.4     8 460      215  3      5.42  17.8     0  0     3     4
```

```
df %>%
  slice_max(order_by = mpg, n = 3)
```

```
## # A tibble: 4 x 12
##   car      mpg   cyl  disp    hp  drat    wt   qsec    vs  am   gear  carb
##   <chr>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <fct> <fct> <fct>
## 1 Toyota Coro~ 33.9     4  71.1    65  4.22  1.84  19.9     1  1     4     1
## 2 Fiat 128      32.4     4  78.7    66  4.08  2.2   19.5     1  1     4     1
## 3 Honda Civic   30.4     4  75.7    52  4.93  1.62  18.5     1  1     4     2
## 4 Lotus Europa  30.4     4  95.1   113  3.77  1.51  16.9     1  1     5     2
```

```
df %>%
  slice_min(order_by = mpg, n = 3)
```

```
## # A tibble: 3 x 12
##   car          mpg   cyl  disp    hp  drat    wt   qsec    vs  am   gear  carb
##   <chr>        <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <fct> <fct> <fct>
## 1 Cadillac Fl~ 10.4     8   472   205  2.93  5.25  18.0     0  0     3     4
## 2 Lincoln Con~ 10.4     8   460   215   3     5.42  17.8     0  0     3     4
## 3 Camaro Z28   13.3     8   350   245  3.73  3.84  15.4     0  0     3     4
```

```
df %>%
  slice_sample(prop = .2, replace = FALSE) # 6/32=0.2
```

```
## # A tibble: 6 x 12
##   car          mpg   cyl  disp    hp  drat    wt   qsec    vs  am   gear  carb
##   <chr>        <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <fct> <fct> <fct>
## 1 Camaro Z28   13.3     8   350   245  3.73  3.84  15.4     0  0     3     4
## 2 Lincoln Con~ 10.4     8   460   215   3     5.42  17.8     0  0     3     4
## 3 Lotus Europa 30.4     4   95.1  113  3.77  1.51  16.9     1  1     5     2
## 4 Merc 280     19.2     6  168.   123  3.92  3.44  18.3     1  0     4     4
## 5 Merc 230     22.8     4  141.    95  3.92  3.15  22.9     1  0     4     2
## 6 AMC Javelin  15.2     8   304   150  3.15  3.44  17.3     0  0     3     2
```

Columns

Select

```
df %>%
  select(car, AM = am, gear, carb, mpg, hp, disp, everything()) %>%
  rename(GEAR = gear)
```

```
## # A tibble: 32 x 12
##   car          AM  GEAR  carb    mpg    hp  disp   cyl  drat    wt   qsec    vs
##   <chr>        <fct> <fct> <fct> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 Mazda RX4    1     4     4     21    110  160     6   3.9   2.62  16.5     0
## 2 Mazda RX4 ~ 1     4     4     21    110  160     6   3.9   2.88  17.0     0
## 3 Datsun 710    1     4     1    22.8    93  108     4   3.85  2.32  18.6     1
## 4 Hornet 4 D~ 0     3     1    21.4   110  258     6   3.08  3.22  19.4     1
## 5 Hornet Spo~ 0     3     2    18.7   175  360     8   3.15  3.44  17.0     0
## 6 Valiant      0     3     1    18.1   105  225     6   2.76  3.46  20.2     1
## 7 Duster 360   0     3     4    14.3   245  360     8   3.21  3.57  15.8     0
## 8 Merc 240D    0     4     2    24.4    62  147.     4   3.69  3.19  20       1
## 9 Merc 230     0     4     2    22.8    95  141.     4   3.92  3.15  22.9     1
## 10 Merc 280    0     4     4    19.2   123  168.     6   3.92  3.44  18.3     1
## # ... with 22 more rows
```

```
df %>%
  select(matches("m|p"))
```

```
## # A tibble: 32 x 4
##       mpg   disp  hp am
##   <dbl> <dbl> <dbl> <fct>
## 1  21     160   110 1
## 2  21     160   110 1
## 3  22.8   108    93 1
## 4  21.4   258   110 0
## 5  18.7   360   175 0
## 6  18.1   225   105 0
## 7  14.3   360   245 0
## 8  24.4   147.    62 0
## 9  22.8   141.    95 0
## 10 19.2   168.   123 0
## # ... with 22 more rows
```

```
# select(!matches("m|p")) # same as below
# select(-matches("m|p"))
```

Mutate

```
df %>%
  mutate(k = rep(1:4, 8))
```

```
## # A tibble: 32 x 13
##       car      mpg   cyl  disp    hp  drat    wt  qsec    vs  am  gear  carb    k
##   <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <fct> <fct> <fct> <int>
## 1 Mazd~  21     6  160   110  3.9   2.62  16.5    0  1    4     4     1
## 2 Mazd~  21     6  160   110  3.9   2.88  17.0    0  1    4     4     2
## 3 Dats~  22.8   4  108    93  3.85  2.32  18.6    1  1    4     1     3
## 4 Horn~  21.4   6  258   110  3.08  3.22  19.4    1  0    3     1     4
## 5 Horn~  18.7   8  360   175  3.15  3.44  17.0    0  0    3     2     1
## 6 Vali~  18.1   6  225   105  2.76  3.46  20.2    1  0    3     1     2
## 7 Dust~  14.3   8  360   245  3.21  3.57  15.8    0  0    3     4     3
## 8 Merc~  24.4   4  147.    62  3.69  3.19  20      1  0    4     2     4
## 9 Merc~  22.8   4  141.    95  3.92  3.15  22.9    1  0    4     2     1
## 10 Merc~ 19.2   6  168.   123  3.92  3.44  18.3    1  0    4     4     2
## # ... with 22 more rows
```

```
df %>%
  mutate(hp1 = hp / 100,
         hp2 = paste0("HP ", hp1),
         hp3 = ifelse(hp > 250, "yes", "no"),
         hp4 = recode(hp3, "yes" = "y", "no" = "n"),
         hp5 = case_when(hp < 160 & am == 1 ~ "low",
                        hp < 250 & am == 0 ~ "medium",
                        hp > 250 ~ "high",
                        TRUE ~ "unknown")) %>%
  select(hp, hp1:last_col()) %>%
  count(hp5)
```



```
## # A tibble: 4 x 2
##   hp5      n
##   <chr>  <int>
## 1 high      2
## 2 low       10
## 3 medium    19
## 4 unknown    1
```

Summarise

```
df %>%
  summarise(across(where(is.numeric), mean))
```

```
## # A tibble: 1 x 8
##   mpg   cyl  disp    hp  drat    wt   qsec    vs
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1  20.1   6.19  231.  147.   3.60   3.22  17.8   0.438
```

```
df %>%
  group_by(am, gear) %>%
  summarise(number = n(),
            across(c(mpg, hp), list(mean = mean, sd = sd)),
            range_hp = range(hp),
            range_between_hp = max(hp) - min(hp),
            first = first(hp),
            .groups = "drop")
```

```
## # A tibble: 8 x 10
##   am   gear number mpg_mean mpg_sd hp_mean hp_sd range_hp range_between_hp
##   <fct> <fct>   <int>   <dbl>   <dbl>   <dbl> <dbl>   <dbl>           <dbl>
## 1  0     3     15    16.1    3.37   176.   47.7     97             148
## 2  0     3     15    16.1    3.37   176.   47.7    245             148
## 3  0     4      4    21.0    3.07   101.   29.0     62              61
## 4  0     4      4    21.0    3.07   101.   29.0    123              61
## 5  1     4      8    26.3    5.41    83.9   24.2     52              58
## 6  1     4      8    26.3    5.41    83.9   24.2    110              58
## 7  1     5      5    21.4    6.66   196.  103.     91             244
## 8  1     5      5    21.4    6.66   196.  103.    335             244
## # ... with 1 more variable: first <dbl>
```

Strings

Test your regex here (<https://regexr.com/>)

1. ^start, \$end, .any, *0 more, ?optional, +,
2. \escape, |or, (group), [one], {times},

```
df %>%  
  .$car %>%  
  nchar()
```

```
## [1] 9 13 10 14 17 7 10 9 8 8 9 10 10 11 18 19 17 8 11 14 13 16 11 10 16  
## [26] 9 13 12 14 12 13 10
```

```
# str_length()
```

```
paste(1:6, "mpg", sep = '|', collapse = " ~ ") # a single string
```

```
## [1] "1|mpg ~ 2|mpg ~ 3|mpg ~ 4|mpg ~ 5|mpg ~ 6|mpg"
```

```
str_c(letters[1:6], 1:6, sep = "**")
```

```
## [1] "a*1" "b*2" "c*3" "d*4" "e*5" "f*6"
```

```
head(df) %>%  
  mutate(text = str_glue("{am}-{gear}-{carb} has horse power of {new_var}", new_var = h  
p)) %>%  
  .$text %>%  
  str_wrap(20) %>%  
  paste0("\n\n") %>% # paragraph***  
  cat()
```

```
## 1-4-4 has horse  
## power of 110  
##  
## 1-4-4 has horse  
## power of 110  
##  
## 1-4-1 has horse  
## power of 93  
##  
## 0-3-1 has horse  
## power of 110  
##  
## 0-3-2 has horse  
## power of 175  
##  
## 0-3-1 has horse  
## power of 105
```

```
df %>%  
  mutate(new = substr(car, 1, 3))
```

```
## # A tibble: 32 x 13
##   car      mpg   cyl  disp    hp  drat    wt   qsec    vs  am   gear  carb  new
##   <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <fct> <fct> <fct> <chr>
## 1 Mazd~    21     6   160   110   3.9   2.62  16.5    0  1     4     4   Maz
## 2 Mazd~    21     6   160   110   3.9   2.88  17.0    0  1     4     4   Maz
## 3 Dats~   22.8     4   108    93   3.85   2.32  18.6    1  1     4     1   Dat
## 4 Horn~   21.4     6   258   110   3.08   3.22  19.4    1  0     3     1   Hor
## 5 Horn~   18.7     8   360   175   3.15   3.44  17.0    0  0     3     2   Hor
## 6 Vali~   18.1     6   225   105   2.76   3.46  20.2    1  0     3     1   Val
## 7 Dust~   14.3     8   360   245   3.21   3.57  15.8    0  0     3     4   Dus
## 8 Merc~   24.4     4   147.    62   3.69   3.19   20     1  0     4     2   Mer
## 9 Merc~   22.8     4   141.    95   3.92   3.15  22.9    1  0     4     2   Mer
## 10 Merc~  19.2     6   168.   123   3.92   3.44  18.3    1  0     4     4   Mer
## # ... with 22 more rows
```

```
# mutate(new = str_sub(car, 1, -3)) # from right
```

```
str_pad("you are so", side = "both", 20)
```

```
## [1] "      you are so      "
```

```
str_trim("      you are      so      ")
```

```
## [1] "you are      so"
```

```
str_squish("      you are      so      ")
```

```
## [1] "you are so"
```

```
str_subset(c("you are so cool", "hello world"), "[h]") # which one contains h
```

```
## [1] "hello world"
```

```
str_which(c("you are so cool", "hello world"), "[h]")
```

```
## [1] 2
```

```
df %>%
  mutate(new = str_split_fixed(car, " ", n = 2)) # better
```

```
## # A tibble: 32 x 13
##   car      mpg   cyl  disp    hp  drat    wt   qsec    vs  am   gear  carb
##   <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <fct> <fct> <fct>
## 1 Mazd~ 21     6  160   110   3.9   2.62  16.5    0  1     4     4
## 2 Mazd~ 21     6  160   110   3.9   2.88  17.0    0  1     4     4
## 3 Dats~ 22.8    4  108    93   3.85   2.32  18.6    1  1     4     1
## 4 Horn~ 21.4    6  258   110   3.08   3.22  19.4    1  0     3     1
## 5 Horn~ 18.7    8  360   175   3.15   3.44  17.0    0  0     3     2
## 6 Vali~ 18.1    6  225   105   2.76   3.46  20.2    1  0     3     1
## 7 Dust~ 14.3    8  360   245   3.21   3.57  15.8    0  0     3     4
## 8 Merc~ 24.4    4  147.    62   3.69   3.19   20     1  0     4     2
## 9 Merc~ 22.8    4  141.    95   3.92   3.15  22.9    1  0     4     2
## 10 Merc~ 19.2    6  168.   123   3.92   3.44  18.3    1  0     4     4
## # ... with 22 more rows, and 2 more variables: new[,1] <chr>, [,2] <chr>
```

```
# mutate(new = str_split(car, "\\d.?") %>% unnest(new) # return a list
```

```
df %>%
  mutate(new = str_extract(car, "(\\d\\d?\\d?[-]?\\d?)"))
```

```
## # A tibble: 32 x 13
##   car      mpg   cyl  disp    hp  drat    wt   qsec    vs  am   gear  carb  new
##   <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <fct> <fct> <fct> <chr>
## 1 Mazd~ 21     6  160   110   3.9   2.62  16.5    0  1     4     4     4
## 2 Mazd~ 21     6  160   110   3.9   2.88  17.0    0  1     4     4     4
## 3 Dats~ 22.8    4  108    93   3.85   2.32  18.6    1  1     4     1    710
## 4 Horn~ 21.4    6  258   110   3.08   3.22  19.4    1  0     3     1     4
## 5 Horn~ 18.7    8  360   175   3.15   3.44  17.0    0  0     3     2    <NA>
## 6 Vali~ 18.1    6  225   105   2.76   3.46  20.2    1  0     3     1    <NA>
## 7 Dust~ 14.3    8  360   245   3.21   3.57  15.8    0  0     3     4    360
## 8 Merc~ 24.4    4  147.    62   3.69   3.19   20     1  0     4     2    240
## 9 Merc~ 22.8    4  141.    95   3.92   3.15  22.9    1  0     4     2    230
## 10 Merc~ 19.2    6  168.   123   3.92   3.44  18.3    1  0     4     4    280
## # ... with 22 more rows
```

```
df %>%
  filter(grepl("\\d", car))
```

```
## # A tibble: 17 x 12
##   car      mpg   cyl  disp    hp  drat    wt   qsec    vs  am  gear  carb
##   <chr>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 Mazda RX4      21     6  160    110   3.9   2.62  16.5     0  1     4     4
## 2 Mazda RX4 ~    21     6  160    110   3.9   2.88  17.0     0  1     4     4
## 3 Datsun 710    22.8     4  108     93   3.85   2.32  18.6     1  1     4     1
## 4 Hornet 4 D~   21.4     6  258    110   3.08   3.22  19.4     1  0     3     1
## 5 Duster 360    14.3     8  360    245   3.21   3.57  15.8     0  0     3     4
## 6 Merc 240D    24.4     4  147.     62   3.69   3.19   20      1  0     4     2
## 7 Merc 230     22.8     4  141.     95   3.92   3.15  22.9     1  0     4     2
## 8 Merc 280     19.2     6  168.    123   3.92   3.44  18.3     1  0     4     4
## 9 Merc 280C    17.8     6  168.    123   3.92   3.44  18.9     1  0     4     4
## 10 Merc 450SE   16.4     8  276.    180   3.07   4.07  17.4     0  0     3     3
## 11 Merc 450SL   17.3     8  276.    180   3.07   3.73  17.6     0  0     3     3
## 12 Merc 450SLC  15.2     8  276.    180   3.07   3.78   18      0  0     3     3
## 13 Fiat 128     32.4     4   78.7     66   4.08   2.2   19.5     1  1     4     1
## 14 Camaro Z28   13.3     8  350    245   3.73   3.84  15.4     0  0     3     4
## 15 Fiat X1-9    27.3     4   79      66   4.08   1.94  18.9     1  1     4     1
## 16 Porsche 91~  26      4  120.     91   4.43   2.14  16.7     0  1     5     2
## 17 Volvo 142E   21.4     4  121    109   4.11   2.78  18.6     1  1     4     2
```

```
# filter(str_detect(car, "\\d"))
```

```
df %>%
  mutate(new = str_remove_all(car, "\\d"))
```

```
## # A tibble: 32 x 13
##   car      mpg   cyl  disp    hp  drat    wt   qsec    vs  am  gear  carb  new
##   <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <chr>
## 1 Mazd~    21     6  160    110   3.9   2.62  16.5     0  1     4     4  "Maz~
## 2 Mazd~    21     6  160    110   3.9   2.88  17.0     0  1     4     4  "Maz~
## 3 Dats~   22.8     4  108     93   3.85   2.32  18.6     1  1     4     1  "Dat~
## 4 Horn~   21.4     6  258    110   3.08   3.22  19.4     1  0     3     1  "Hor~
## 5 Horn~   18.7     8  360    175   3.15   3.44  17.0     0  0     3     2  "Hor~
## 6 Vali~   18.1     6  225    105   2.76   3.46  20.2     1  0     3     1  "Val~
## 7 Dust~   14.3     8  360    245   3.21   3.57  15.8     0  0     3     4  "Dus~
## 8 Merc~   24.4     4  147.     62   3.69   3.19   20      1  0     4     2  "Mer~
## 9 Merc~   22.8     4  141.     95   3.92   3.15  22.9     1  0     4     2  "Mer~
## 10 Merc~  19.2     6  168.    123   3.92   3.44  18.3     1  0     4     4  "Mer~
## # ... with 22 more rows
```

```
# mutate(new = str_replace_all(car, "\\d", "A"))
```