Veri Manipulasyonu

mcyflights 2013 yili New york ucus bilgileri

http://www.transtats.bts.gov/DatabaseInfo.asp?DB_ID=120&Link=0

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(nycflights13)

df <- flights

df
```

```
## # A tibble: 336,776 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
##  1  2013     1     1      517            515         2      830            819
##  2  2013     1     1      533            529         4      850            830
##  3  2013     1     1      542            540         2      923            850
##  4  2013     1     1      544            545        -1     1004           1022
```

```
##  5  2013     1     1      554            600        -6      812            837
##  6  2013     1     1      554            558        -4      740            728
##  7  2013     1     1      555            600        -5      913            854
##  8  2013     1     1      557            600        -3      709            723
##  9  2013     1     1      557            600        -3      838            846
## 10  2013     1     1      558            600        -2      753            745
## # ... with 336,766 more rows, and 11 more variables: arr_delay <dbl>,
## #   carrier <chr>, flight <int>, tailnum <chr>, origin <chr>, dest <chr>,
## #   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dttm>
```

```
str(df)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    336776 obs. of  19 variables:
##  $ year          : int  2013 2013 2013 2013 2013 2013 2013 2013 2013 2013 ...
##  $ month         : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ day           : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ dep_time      : int  517 533 542 544 554 554 555 557 557 558 ...
##  $ sched_dep_time: int  515 529 540 545 600 558 600 600 600 600 ...
##  $ dep_delay     : num  2 4 2 -1 -6 -4 -5 -3 -3 -2 ...
##  $ arr_time      : int  830 850 923 1004 812 740 913 709 838 753 ...
##  $ sched_arr_time: int  819 830 850 1022 837 728 854 723 846 745 ...
##  $ arr_delay     : num  11 20 33 -18 -25 12 19 -14 -8 8 ...
##  $ carrier       : chr  "UA" "UA" "AA" "B6" ...
##  $ flight        : int  1545 1714 1141 725 461 1696 507 5708 79 301 ...
##  $ tailnum       : chr  "N14228" "N24211" "N619AA" "N804JB" ...
##  $ origin        : chr  "EWR" "LGA" "JFK" "JFK" ...
##  $ dest          : chr  "IAH" "IAH" "MIA" "BQN" ...
##  $ air_time      : num  227 227 160 183 116 150 158 53 140 138 ...
##  $ distance      : num  1400 1416 1089 1576 762 ...
##  $ hour          : num  5 5 5 5 6 5 6 6 6 6 ...
##  $ minute        : num  15 29 40 45 0 58 0 0 0 0 ...
##  $ time_hour     : POSIXct, format: "2013-01-01 05:00:00" "2013-01-01 05:00:00" ...
```

```
summary(df)
```

```
##       year          month             day           dep_time    sched_dep_time
##  Min.   :2013   Min.   : 1.000   Min.   : 1.00   Min.   :   1   Min.   : 106
##  1st Qu.:2013   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.: 907   1st Qu.: 906
##  Median :2013   Median : 7.000   Median :16.00   Median :1401   Median :1359
##  Mean   :2013   Mean   : 6.549   Mean   :15.71   Mean   :1349   Mean   :1344
##  3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:1744   3rd Qu.:1729
##  Max.   :2013   Max.   :12.000   Max.   :31.00   Max.   :2400   Max.   :2359
##                                                  NA's   :8255
##    dep_delay          arr_time    sched_arr_time   arr_delay
##  Min.   : -43.00   Min.   :   1   Min.   :   1   Min.   : -86.000
##  1st Qu.:  -5.00   1st Qu.:1104   1st Qu.:1124   1st Qu.: -17.000
##  Median :  -2.00   Median :1535   Median :1556   Median :  -5.000
##  Mean   :  12.64   Mean   :1502   Mean   :1536   Mean   :   6.895
##  3rd Qu.:  11.00   3rd Qu.:1940   3rd Qu.:1945   3rd Qu.:  14.000
##  Max.   :1301.00   Max.   :2400   Max.   :2359   Max.   :1272.000
##  NA's   :8255      NA's   :8713                  NA's   :9430
##    carrier              flight       tailnum             origin
##  Length:336776      Min.   :   1   Length:336776      Length:336776
##  Class :character   1st Qu.: 553   Class :character   Class :character
##  Mode  :character   Median :1496   Mode  :character   Mode  :character
##                     Mean   :1972
##                     3rd Qu.:3465
##                     Max.   :8500
##
##     dest              air_time        distance         hour
##  Length:336776      Min.   : 20.0   Min.   :  17   Min.   : 1.00
##  Class :character   1st Qu.: 82.0   1st Qu.: 502   1st Qu.: 9.00
##  Mode  :character   Median :129.0   Median : 872   Median :13.00
##                     Mean   :150.7   Mean   :1040   Mean   :13.18
##                     3rd Qu.:192.0   3rd Qu.:1389   3rd Qu.:17.00
##                     Max.   :695.0   Max.   :4983   Max.   :23.00
##                     NA's   :9430
##     minute        time_hour
##  Min.   : 0.00   Min.   :2013-01-01 05:00:00
##  1st Qu.: 8.00   1st Qu.:2013-04-04 13:00:00
##  Median :29.00   Median :2013-07-03 10:00:00
##  Mean   :26.23   Mean   :2013-07-03 05:22:54
```

```
##  3rd Qu.:44.00   3rd Qu.:2013-10-01 07:00:00
##  Max.   :59.00   Max.   :2013-12-31 23:00:00
##
```

```r
# select fonksiyonu df icindeki belirtilen degiskenleri secerek bir tablo olusturur
# filter fonksiyonu belirtilen degiskenin belitilen kosuldaki degiskenlerini secer
# group_by donksiyonu belirlenen degiskene gore tabloyu gruplaa ayirir
# summarise ortalama bir deger yapar


df %>%
  select(dep_delay, day, carrier) %>%
  filter(dep_delay > 10) %>%
  group_by(carrier) %>%
  summarise(n = n(),
            ortalama = mean(dep_delay))
```

```
## # A tibble: 16 x 3
##    carrier     n ortalama
##    <chr>   <int>    <dbl>
##  1 9E       5064     66.3
##  2 AA       6012     59.9
##  3 AS        118     56.3
##  4 B6      14454     56.7
##  5 DL       9346     58.1
##  6 EV      17530     64.9
##  7 F9        233     63.7
##  8 FL       1007     64.1
##  9 HA         28    106.
## 10 MQ       6084     57.7
## 11 OO          7     73.9
## 12 UA      14873     51.2
## 13 US       2923     51.1
## 14 VX       1087     66.0
## 15 WN       3890     55.8
## 16 YV        178     67.8
```

degisken islemleri : select()

```
select(df, carrier, flight, tailnum)  # belirtilenleri secme
```

```
## # A tibble: 336,776 x 3
##    carrier flight tailnum
##    <chr>    <int> <chr>
##  1 UA        1545 N14228
##  2 UA        1714 N24211
##  3 AA        1141 N619AA
##  4 B6         725 N804JB
##  5 DL         461 N668DN
##  6 UA        1696 N39463
##  7 B6         507 N516JB
##  8 EV        5708 N829AS
##  9 B6          79 N593JB
## 10 AA         301 N3ALAA
## # ... with 336,766 more rows
```

```
select(df, carrier, origin:hour)  # aralık vererek secme
```

```
## # A tibble: 336,776 x 6
##    carrier origin dest  air_time distance  hour
##    <chr>   <chr>  <chr>    <dbl>    <dbl> <dbl>
##  1 UA      EWR    IAH        227     1400     5
##  2 UA      LGA    IAH        227     1416     5
##  3 AA      JFK    MIA        160     1089     5
##  4 B6      JFK    BQN        183     1576     5
##  5 DL      LGA    ATL        116      762     6
##  6 UA      EWR    ORD        150      719     5
##  7 B6      EWR    FLL        158     1065     6
##  8 EV      LGA    IAD         53      229     6
##  9 B6      JFK    MCO        140      944     6
```

```
## 10 AA      LGA    ORD        138      733     6
## # ... with 336,766 more rows
```

```
select(df, 1:4)  # aralık vererek secme
```

```
## # A tibble: 336,776 x 4
##     year month   day dep_time
##    <int> <int> <int>    <int>
##  1  2013     1     1      517
##  2  2013     1     1      533
##  3  2013     1     1      542
##  4  2013     1     1      544
##  5  2013     1     1      554
##  6  2013     1     1      554
##  7  2013     1     1      555
##  8  2013     1     1      557
##  9  2013     1     1      557
## 10  2013     1     1      558
## # ... with 336,766 more rows
```

```
select(df, -c(carrier, origin))  # belirtilenleri disarda birakir
```

```
## # A tibble: 336,776 x 17
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
##  1  2013     1     1      517            515         2      830            819
##  2  2013     1     1      533            529         4      850            830
##  3  2013     1     1      542            540         2      923            850
##  4  2013     1     1      544            545        -1     1004           1022
##  5  2013     1     1      554            600        -6      812            837
##  6  2013     1     1      554            558        -4      740            728
##  7  2013     1     1      555            600        -5      913            854
##  8  2013     1     1      557            600        -3      709            723
##  9  2013     1     1      557            600        -3      838            846
```

```
## 10  2013     1     1      558            600        -2      753            745
## # ... with 336,766 more rows, and 9 more variables: arr_delay <dbl>,
## #   flight <int>, tailnum <chr>, dest <chr>, air_time <dbl>, distance <dbl>,
## #   hour <dbl>, minute <dbl>, time_hour <dttm>
```

```
select(df, contains("time"))  # belirtilen degiskeni iceren elemanların tablosu
```

```
## # A tibble: 336,776 x 6
##    dep_time sched_dep_time arr_time sched_arr_time air_time time_hour
##       <int>          <int>    <int>          <int>    <dbl> <dttm>
##  1      517            515      830            819      227 2013-01-01 05:00:00
##  2      533            529      850            830      227 2013-01-01 05:00:00
##  3      542            540      923            850      160 2013-01-01 05:00:00
##  4      544            545     1004           1022      183 2013-01-01 05:00:00
##  5      554            600      812            837      116 2013-01-01 06:00:00
##  6      554            558      740            728      150 2013-01-01 05:00:00
##  7      555            600      913            854      158 2013-01-01 06:00:00
##  8      557            600      709            723       53 2013-01-01 06:00:00
##  9      557            600      838            846      140 2013-01-01 06:00:00
## 10      558            600      753            745      138 2013-01-01 06:00:00
## # ... with 336,766 more rows
```

```
select(df, starts_with("dep"))  #  belirtilen degiske ile baslayan elemanların tablosu
```

```
## # A tibble: 336,776 x 2
##    dep_time dep_delay
##       <int>     <dbl>
##  1      517         2
##  2      533         4
##  3      542         2
##  4      544        -1
##  5      554        -6
##  6      554        -4
##  7      555        -5
```

```
##  8      557        -3
##  9      557        -3
## 10      558        -2
## # ... with 336,766 more rows
```

```
select(df, ends_with("delay"))  #  belirtilen degiske ile biten elemanların tablosu
```

```
## # A tibble: 336,776 x 2
##    dep_delay arr_delay
##        <dbl>     <dbl>
##  1         2        11
##  2         4        20
##  3         2        33
##  4        -1       -18
##  5        -6       -25
##  6        -4        12
##  7        -5        19
##  8        -3       -14
##  9        -3        -8
## 10        -2         8
## # ... with 336,766 more rows
```

```
m <- matrix(1:25, 5, 5 )

colnames(m) <- paste("x", 1:5, sep = "")

select(as.data.frame(m), num_range("x", 1:3)) # belirtilen data frame icinde x in yanında farkli degerler olunca
 ise yarar
```

```
##   x1 x2 x3
## 1  1  6 11
## 2  2  7 12
## 3  3  8 13
```

```
## 4  4  9 14
## 5  5 10 15
```

```
select(df, carrier, tailnum, contains("time"))  #farkli kullnaimlari da mevcuttur
```

```
## # A tibble: 336,776 x 8
##    carrier tailnum dep_time sched_dep_time arr_time sched_arr_time air_time
##    <chr>   <chr>      <int>          <int>    <int>          <int>    <dbl>
##  1 UA      N14228       517            515      830            819      227
##  2 UA      N24211       533            529      850            830      227
##  3 AA      N619AA       542            540      923            850      160
##  4 B6      N804JB       544            545     1004           1022      183
##  5 DL      N668DN       554            600      812            837      116
##  6 UA      N39463       554            558      740            728      150
##  7 B6      N516JB       555            600      913            854      158
##  8 EV      N829AS       557            600      709            723       53
##  9 B6      N593JB       557            600      838            846      140
## 10 AA      N3ALAA       558            600      753            745      138
## # ... with 336,766 more rows, and 1 more variable: time_hour <dttm>
```

Gozlem islemleri : filter()

```
filter(df, year == 2013 & month == 2) # belitilen durumlara gore filtreleyipo bize verir
```

```
## # A tibble: 24,951 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
##  1  2013     2     1      456            500        -4      652            648
##  2  2013     2     1      520            525        -5      816            820
##  3  2013     2     1      527            530        -3      837            829
##  4  2013     2     1      532            540        -8     1007           1017
##  5  2013     2     1      540            540         0      859            850
##  6  2013     2     1      552            600        -8      714            715
```

```
##  7  2013     2     1      552            600        -8      919            910
##  8  2013     2     1      552            600        -8      655            709
##  9  2013     2     1      553            600        -7      833            815
## 10  2013     2     1      553            600        -7      821            825
## # ... with 24,941 more rows, and 11 more variables: arr_delay <dbl>,
## #   carrier <chr>, flight <int>, tailnum <chr>, origin <chr>, dest <chr>,
## #   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dttm>
```

```
filter(df, dep_delay > mean(df$dep_delay, na.rm = TRUE) + sd(df$dep_delay, na.rm = TRUE))  # gecikmesi ortlama ge
cikmeden kucuk olanlar ve standartdan sapmasından
```

```
## # A tibble: 30,888 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
##  1  2013     1     1      811            630       101     1047            830
##  2  2013     1     1      826            715        71     1136           1045
##  3  2013     1     1      848           1835       853     1001           1950
##  4  2013     1     1      909            810        59     1331           1315
##  5  2013     1     1      957            733       144     1056            853
##  6  2013     1     1     1114            900       134     1447           1222
##  7  2013     1     1     1120            944        96     1331           1213
##  8  2013     1     1     1255           1200        55     1451           1330
##  9  2013     1     1     1301           1150        71     1518           1345
## 10  2013     1     1     1337           1220        77     1649           1531
## # ... with 30,878 more rows, and 11 more variables: arr_delay <dbl>,
## #   carrier <chr>, flight <int>, tailnum <chr>, origin <chr>, dest <chr>,
## #   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dttm>
```

```
df %>%
  filter(month == 2, day == 18) %>%
  select(dep_delay, month) %>%
  arrange(desc(dep_delay), month)  # belirtilen degiskenlere gore siralar
```

```
## # A tibble: 948 x 2
##    dep_delay month
##        <dbl> <int>
##  1       281     2
##  2       247     2
##  3       221     2
##  4       216     2
##  5       208     2
##  6       204     2
##  7       195     2
##  8       174     2
##  9       154     2
## 10       152     2
## # ... with 938 more rows
```

```
df %>% sample_frac(0.05) #belirli bir oranda rastgele secim yapiyor
```

```
## # A tibble: 16,839 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
##  1  2013    10    14     1916           1900        16     2102           2117
##  2  2013     3    17     1802           1750        12     2126           2105
##  3  2013     5    19     1144           1140         4     1243           1247
##  4  2013     7    13     1259           1255         4     1553           1605
##  5  2013     5    19     2013           1940        33     2252           2252
##  6  2013    12    15     2045           2025        20     2203           2205
##  7  2013     5    29     1314           1315        -1     1533           1538
##  8  2013     1     3     2019           2000        19     2303           2325
##  9  2013     4     4      957            915        42     1326           1316
## 10  2013     7    13      628            630        -2      800            804
## # ... with 16,829 more rows, and 11 more variables: arr_delay <dbl>,
## #   carrier <chr>, flight <int>, tailnum <chr>, origin <chr>, dest <chr>,
## #   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dttm>
```

```
df %>% sample_n(20) #belirli bir degerde rastgele secim yapiyor
```

```
## # A tibble: 20 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
##  1  2013    11     5     1021           1021         0     1314           1325
##  2  2013     2     8      927            815        72     1139           1010
##  3  2013    12    27      902            835        27     1042           1040
##  4  2013    12    15     1826           1830        -4     2047           2054
##  5  2013    12    15     1104           1105        -1     1235           1242
##  6  2013     4    21     1457           1459        -2     1810           1801
##  7  2013     7    25     2128           2108        20       10           2359
##  8  2013     2     7     1350           1300        50     1526           1440
##  9  2013     8    25      645            641         4      803            807
## 10  2013     1    13     1739           1724        15     2013           1953
## 11  2013     3     6     1053           1010        43     1200           1125
## 12  2013     3     3     1716           1715         1     2021           2027
## 13  2013     8    22     1629           1625         4     1822           1834
## 14  2013    11     5     1215           1219        -4     1407           1416
## 15  2013     5    10     1633           1640        -7     1853           1845
## 16  2013    12    12     1902           1900         2     2056           2057
## 17  2013     4     3     1607           1600         7     1718           1720
## 18  2013    12     2      600            600         0      750            810
## 19  2013    10    17     1615           1555        20     1829           1755
## 20  2013    10    16      614            620        -6      714            728
## # ... with 11 more variables: arr_delay <dbl>, carrier <chr>, flight <int>,
## #   tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>,
## #   hour <dbl>, minute <dbl>, time_hour <dttm>
```

```
df %>% slice(1:20)  #pozisyona gore secim yapiyor
```

```
## # A tibble: 20 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
```

```
##  1  2013     1     1      517            515         2      830            819
##  2  2013     1     1      533            529         4      850            830
##  3  2013     1     1      542            540         2      923            850
##  4  2013     1     1      544            545        -1     1004           1022
##  5  2013     1     1      554            600        -6      812            837
##  6  2013     1     1      554            558        -4      740            728
##  7  2013     1     1      555            600        -5      913            854
##  8  2013     1     1      557            600        -3      709            723
##  9  2013     1     1      557            600        -3      838            846
## 10  2013     1     1      558            600        -2      753            745
## 11  2013     1     1      558            600        -2      849            851
## 12  2013     1     1      558            600        -2      853            856
## 13  2013     1     1      558            600        -2      924            917
## 14  2013     1     1      558            600        -2      923            937
## 15  2013     1     1      559            600        -1      941            910
## 16  2013     1     1      559            559         0      702            706
## 17  2013     1     1      559            600        -1      854            902
## 18  2013     1     1      600            600         0      851            858
## 19  2013     1     1      600            600         0      837            825
## 20  2013     1     1      601            600         1      844            850
## # ... with 11 more variables: arr_delay <dbl>, carrier <chr>, flight <int>,
## #   tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>,
## #   hour <dbl>, minute <dbl>, time_hour <dttm>
```

```
df %>% top_n(10)  #ilk n degiskeni secer
```

```
## Selecting by time_hour
```

```
## # A tibble: 12 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
##  1  2013    12    31       13           2359        14      439            437
##  2  2013    12    31       18           2359        19      449            444
##  3  2013    12    31       26           2245       101      129           2353
##  4  2013    12    31     2218           2219        -1      315            304
```

```
##  5  2013    12    31     2235           2245       -10     2351           2355
##  6  2013    12    31     2245           2250        -5     2359           2356
##  7  2013    12    31     2310           2255        15        7           2356
##  8  2013    12    31     2321           2250        31       46              8
##  9  2013    12    31     2328           2330        -2      412            409
## 10  2013    12    31     2332           2245        47       58              3
## 11  2013    12    31     2355           2359        -4      430            440
## 12  2013    12    31     2356           2359        -3      436            445
## # ... with 11 more variables: arr_delay <dbl>, carrier <chr>, flight <int>,
## #   tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>,
## #   hour <dbl>, minute <dbl>, time_hour <dttm>
```

Degisken donusum islemleri : mutate()

```
sample <- df %>%
  sample_n(20) %>%
  select(arr_delay, dep_delay, distance, arr_time)

sample
```

```
## # A tibble: 20 x 4
##    arr_delay dep_delay distance arr_time
##        <dbl>     <dbl>    <dbl>    <int>
##  1        10        15      719     1849
##  2        80        49      200     2213
##  3        15         3     1065     1209
##  4       -14        -5     1028     1150
##  5       -16        -7     1183     1612
##  6       -26        -5      187     2123
##  7        99       100     2586     1554
##  8         2         8      541     1914
##  9         8        34     2133     1623
## 10        23        14      760      906
## 11       354       375      301      437
## 12         0         6     1391     1930
```

```
## 13       -24        -3     1725     1158
## 14        17        -5      431     1947
## 15        14        -6      431     1404
## 16       -12        12     2475     1412
## 17        15        21      544     1725
## 18       -31        -7      187     1246
## 19        34         0     1372     2215
## 20        18        -2      719      918
```

```
m <- mutate(sample,
       kazanc = arr_delay - dep_delay,
       hiz = distance / arr_time * 60,
       yeni = kazanc / hiz) #yeni degisken olusturulur


rename(m,
       kaz = kazanc,
       h = hiz)  # isimmlendirme yapar
```

```
## # A tibble: 20 x 7
##    arr_delay dep_delay distance arr_time   kaz      h    yeni
##        <dbl>     <dbl>    <dbl>    <int> <dbl>  <dbl>   <dbl>
##  1        10        15      719     1849    -5  23.3  -0.214
##  2        80        49      200     2213    31   5.42  5.72
##  3        15         3     1065     1209    12  52.9   0.227
##  4       -14        -5     1028     1150    -9  53.6  -0.168
##  5       -16        -7     1183     1612    -9  44.0  -0.204
##  6       -26        -5      187     2123   -21   5.28 -3.97
##  7        99       100     2586     1554    -1  99.8  -0.0100
##  8         2         8      541     1914    -6  17.0  -0.354
##  9         8        34     2133     1623   -26  78.9  -0.330
## 10        23        14      760      906     9  50.3   0.179
## 11       354       375      301      437   -21  41.3  -0.508
## 12         0         6     1391     1930    -6  43.2  -0.139
## 13       -24        -3     1725     1158   -21  89.4  -0.235
## 14        17        -5      431     1947    22  13.3   1.66
```

```
## 15        14        -6      431     1404    20  18.4   1.09
## 16       -12        12     2475     1412   -24 105.   -0.228
## 17        15        21      544     1725    -6  18.9  -0.317
## 18       -31        -7      187     1246   -24   9.00 -2.67
## 19        34         0     1372     2215    34  37.2   0.915
## 20        18        -2      719      918    20  47.0   0.426
```

```
transmute(sample,
       kazanc = arr_delay - dep_delay,
       hiz = distance / arr_time * 60,
       yeni = kazanc / hiz) # yeni degisken olusturdaktan sonra eskileri istemezsek bu fonksiyon bunu saglar
```

```
## # A tibble: 20 x 3
##    kazanc    hiz    yeni
##     <dbl>  <dbl>   <dbl>
##  1     -5  23.3  -0.214
##  2     31   5.42  5.72
##  3     12  52.9   0.227
##  4     -9  53.6  -0.168
##  5     -9  44.0  -0.204
##  6    -21   5.28 -3.97
##  7     -1  99.8  -0.0100
##  8     -6  17.0  -0.354
##  9    -26  78.9  -0.330
## 10      9  50.3   0.179
## 11    -21  41.3  -0.508
## 12     -6  43.2  -0.139
## 13    -21  89.4  -0.235
## 14     22  13.3   1.66
## 15     20  18.4   1.09
## 16    -24 105.   -0.228
## 17     -6  18.9  -0.317
## 18    -24   9.00 -2.67
## 19     34  37.2   0.915
## 20     20  47.0   0.426
```

Gruplama ve veri ozetleme: group_by()

```
# group_by fonksiyonu belirlenen deiskene gore gruplar

df %>% sample_n(20) %>%
  group_by(carrier) %>%
  summarise(sayi = n(),
            ortalama = mean(dep_delay, na.rm = TRUE),
            medyan = median(dep_delay, na.rm = TRUE),
            sd = sd(dep_delay, na.rm = TRUE))
```

```
## # A tibble: 9 x 5
##   carrier  sayi ortalama medyan     sd
##   <chr>   <int>    <dbl>  <dbl>  <dbl>
## 1 9E          2   -10     -10   NaN
## 2 AA          2    -2      -2     2.83
## 3 B6          4     7.75    5    11.8
## 4 DL          2     0.5     0.5   2.12
## 5 EV          2    -6      -6   NaN
## 6 MQ          1   NaN      NA   NaN
## 7 UA          4    11.5    10.5  16.8
## 8 US          2    -4.5    -4.5   2.12
## 9 WN          1    73      73   NaN
```

Tidy data gathering ve spreading : select()

```
library(tidyverse)
```

```
## -- Attaching packages --------
```

```
## <U+221A> ggplot2 3.2.1     <U+221A> purrr   0.3.3
## <U+221A> tibble  2.1.3     <U+221A> stringr 1.4.0
```

```
## <U+221A> tidyr   1.0.2     <U+221A> forcats 0.4.0
## <U+221A> readr   1.3.1
```

```
## -- Conflicts -----------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
# kutuphanenin icindeki tablolar

#tablolarda belirli bozukluklar var bunlar düzeltilecek

table1
```

```
## # A tibble: 6 x 4
##   country      year  cases population
##   <chr>       <int>  <int>      <int>
## 1 Afghanistan  1999    745   19987071
## 2 Afghanistan  2000   2666   20595360
## 3 Brazil       1999  37737  172006362
## 4 Brazil       2000  80488  174504898
## 5 China        1999 212258 1272915272
## 6 China        2000 213766 1280428583
```

```
table2
```

```
## # A tibble: 12 x 4
##    country      year type            count
##    <chr>       <int> <chr>           <int>
##  1 Afghanistan  1999 cases             745
##  2 Afghanistan  1999 population   19987071
##  3 Afghanistan  2000 cases            2666
##  4 Afghanistan  2000 population   20595360
##  5 Brazil       1999 cases           37737
##  6 Brazil       1999 population  172006362
```

```
##  7 Brazil       2000 cases           80488
##  8 Brazil       2000 population  174504898
##  9 China        1999 cases          212258
## 10 China        1999 population 1272915272
## 11 China        2000 cases          213766
## 12 China        2000 population 1280428583
```

```
table3
```

```
## # A tibble: 6 x 3
##   country      year rate
## * <chr>       <int> <chr>
## 1 Afghanistan  1999 745/19987071
## 2 Afghanistan  2000 2666/20595360
## 3 Brazil       1999 37737/172006362
## 4 Brazil       2000 80488/174504898
## 5 China        1999 212258/1272915272
## 6 China        2000 213766/1280428583
```

```
table4a
```

```
## # A tibble: 3 x 3
##   country     `1999` `2000`
## * <chr>        <int>  <int>
## 1 Afghanistan    745   2666
## 2 Brazil       37737  80488
## 3 China       212258 213766
```

```
table4b
```

```
## # A tibble: 3 x 3
##   country         `1999`     `2000`
## * <chr>            <int>      <int>
```

```
## 1 Afghanistan   19987071   20595360
## 2 Brazil       172006362  174504898
## 3 China       1272915272 1280428583
```

```
# gather fonksiyonu getirmek istedigimiz degiskenleri tek bir satira toplar

# bozuk yapidaki tablonun 1999 ve 2000 adli stunlarini year stunu altina toplayip degerlerini de cases stuunu alt
ina topladik

tidya <- table4a %>% gather('1999', '2000', key = "year", value = "cases")

tidya
```

```
## # A tibble: 6 x 3
##   country     year   cases
##   <chr>       <chr>  <int>
## 1 Afghanistan 1999     745
## 2 Brazil      1999   37737
## 3 China       1999  212258
## 4 Afghanistan 2000    2666
## 5 Brazil      2000   80488
## 6 China       2000  213766
```

```
# ayni durumdaki table4b de duzenlendi

tidyb <- table4b %>% gather('1999', '2000', key = "year", value = "population")

tidyb
```

```
## # A tibble: 6 x 3
##   country     year  population
##   <chr>       <chr>      <int>
## 1 Afghanistan 1999    19987071
## 2 Brazil      1999   172006362
```

```
## 3 China       1999  1272915272
## 4 Afghanistan 2000    20595360
## 5 Brazil      2000   174504898
## 6 China       2000  1280428583
```

```
# bu iki tabloyu tek bir tab lo olarak olusturucaz

left_join(tidya, tidyb)
```

```
## Joining, by = c("country", "year")
```

```
## # A tibble: 6 x 4
##   country     year   cases population
##   <chr>       <chr>  <int>      <int>
## 1 Afghanistan 1999     745   19987071
## 2 Brazil      1999   37737  172006362
## 3 China       1999  212258 1272915272
## 4 Afghanistan 2000    2666   20595360
## 5 Brazil      2000   80488  174504898
## 6 China       2000  213766 1280428583
```

```
#spreaing fonksiyonu gather fonksiyonun tersidir

table2
```

```
## # A tibble: 12 x 4
##    country      year type            count
##    <chr>       <int> <chr>           <int>
##  1 Afghanistan  1999 cases             745
##  2 Afghanistan  1999 population   19987071
##  3 Afghanistan  2000 cases            2666
##  4 Afghanistan  2000 population   20595360
##  5 Brazil       1999 cases           37737
##  6 Brazil       1999 population  172006362
```

```
##  7 Brazil       2000 cases           80488
##  8 Brazil       2000 population  174504898
##  9 China        1999 cases          212258
## 10 China        1999 population 1272915272
## 11 China        2000 cases          213766
## 12 China        2000 population 1280428583
```

```
# table2 deki cases sutunun altindaki cases ve population degiskenlerinden yeni suunlar olusturup altlarina da co
unt degerlerini yaziyoruz
spread(table2, key = "type", value = "count")
```

```
## # A tibble: 6 x 4
##   country      year  cases population
##   <chr>       <int>  <int>      <int>
## 1 Afghanistan  1999    745   19987071
## 2 Afghanistan  2000   2666   20595360
## 3 Brazil       1999  37737  172006362
## 4 Brazil       2000  80488  174504898
## 5 China        1999 212258 1272915272
## 6 China        2000 213766 1280428583
```

```
#separeting fonksiyonu bir degiskenin degerini iki degiskenin degeri olarak donusturur

table3
```

```
## # A tibble: 6 x 3
##   country      year rate
## * <chr>       <int> <chr>
## 1 Afghanistan  1999 745/19987071
## 2 Afghanistan  2000 2666/20595360
## 3 Brazil       1999 37737/172006362
## 4 Brazil       2000 80488/174504898
## 5 China        1999 212258/1272915272
## 6 China        2000 213766/1280428583
```

```r
# table3 de rate degikenin altindaki degerleri cases ve poopulation adli iki ayri degiskene bolucez

table3 %>% separate(rate, into = c("cases", "population"), sep = "/", convert = TRUE)
```

```
## # A tibble: 6 x 4
##   country      year  cases population
##   <chr>       <int>  <int>      <int>
## 1 Afghanistan  1999    745   19987071
## 2 Afghanistan  2000   2666   20595360
## 3 Brazil       1999  37737  172006362
## 4 Brazil       2000  80488  174504898
## 5 China        1999 212258 1272915272
## 6 China        2000 213766 1280428583
```

```r
#unite fonksiyonu separeting fonksiyonun tersidir bu da iki degiskeni tek degisken altinda toplar

table5
```

```
## # A tibble: 6 x 4
##   country     century year  rate
## * <chr>       <chr>   <chr> <chr>
## 1 Afghanistan 19      99    745/19987071
## 2 Afghanistan 20      00    2666/20595360
## 3 Brazil      19      99    37737/172006362
## 4 Brazil      20      00    80488/174504898
## 5 China       19      99    212258/1272915272
## 6 China       20      00    213766/1280428583
```

```r
# table5 deki century ve year degsikenlerini tek degisken olarak yazdik

table5 %>% unite(new, century, year, sep = "")
```

```
## # A tibble: 6 x 3
##   country     new   rate
##   <chr>       <chr> <chr>
## 1 Afghanistan 1999  745/19987071
## 2 Afghanistan 2000  2666/20595360
## 3 Brazil      1999  37737/172006362
## 4 Brazil      2000  80488/174504898
## 5 China       1999  212258/1272915272
## 6 China       2000  213766/1280428583
```