31 Oct

Ananya

2023-10-31

Mutate command: to add new variables

syntax: mutate(data,new colname…define the col)

```
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
```

```
rm(iris) # to remove the data
```

```
## Warning in rm(iris): object 'iris' not found
```

```
iris %>% head()
```

```
##   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
```

```
iris=mutate(iris,Sepal.Mean=mean(c(Sepal.Length,Sepal.Width))) %>% head() # it will show 6 co
lumns and we are storing this data into iris


iris
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species Sepal.Mean
## 1          5.1         3.5          1.4         0.2  setosa   4.450333
## 2          4.9         3.0          1.4         0.2  setosa   4.450333
## 3          4.7         3.2          1.3         0.2  setosa   4.450333
## 4          4.6         3.1          1.5         0.2  setosa   4.450333
## 5          5.0         3.6          1.4         0.2  setosa   4.450333
## 6          5.4         3.9          1.7         0.4  setosa   4.450333
```

```
library(hflights)
hflights %>% head()
```

```
##      Year Month DayofMonth DayOfWeek DepTime ArrTime UniqueCarrier FlightNum
## 5424 2011     1          1         6    1400    1500            AA       428
## 5425 2011     1          2         7    1401    1501            AA       428
## 5426 2011     1          3         1    1352    1502            AA       428
## 5427 2011     1          4         2    1403    1513            AA       428
## 5428 2011     1          5         3    1405    1507            AA       428
## 5429 2011     1          6         4    1359    1503            AA       428
##      TailNum ActualElapsedTime AirTime ArrDelay DepDelay Origin Dest Distance
## 5424  N576AA                60      40      -10        0    IAH  DFW      224
## 5425  N557AA                60      45       -9        1    IAH  DFW      224
## 5426  N541AA                70      48       -8       -8    IAH  DFW      224
## 5427  N403AA                70      39        3        3    IAH  DFW      224
## 5428  N492AA                62      44       -3        5    IAH  DFW      224
## 5429  N262AA                64      45       -7       -1    IAH  DFW      224
##      TaxiIn TaxiOut Cancelled CancellationCode Diverted
## 5424      7      13         0                         0
## 5425      6       9         0                         0
## 5426      5      17         0                         0
## 5427      9      22         0                         0
## 5428      9       9         0                         0
## 5429      6      13         0                         0
```

```
hflights=hflights %>% mutate(Speed=Distance/AirTime) %>% head()
```

```
##summarise to reduce the variables to values
# primarily useful with data that has been grouped by one or more variables .
# group.by() creates the groups that will be operated on.
# summarise() uses the provided aggregation function to summarise each group

head(hflights)
```

```
##      Year Month DayofMonth DayOfWeek DepTime ArrTime UniqueCarrier FlightNum
## 5424 2011     1          1         6    1400    1500            AA       428
## 5425 2011     1          2         7    1401    1501            AA       428
## 5426 2011     1          3         1    1352    1502            AA       428
## 5427 2011     1          4         2    1403    1513            AA       428
## 5428 2011     1          5         3    1405    1507            AA       428
## 5429 2011     1          6         4    1359    1503            AA       428
##      TailNum ActualElapsedTime AirTime ArrDelay DepDelay Origin Dest Distance
## 5424  N576AA                60      40      -10        0    IAH  DFW      224
## 5425  N557AA                60      45       -9        1    IAH  DFW      224
## 5426  N541AA                70      48       -8       -8    IAH  DFW      224
## 5427  N403AA                70      39        3        3    IAH  DFW      224
## 5428  N492AA                62      44       -3        5    IAH  DFW      224
## 5429  N262AA                64      45       -7       -1    IAH  DFW      224
##      TaxiIn TaxiOut Cancelled CancellationCode Diverted    Speed
## 5424      7      13         0                        0 5.600000
## 5425      6       9         0                        0 4.977778
## 5426      5      17         0                        0 4.666667
## 5427      9      22         0                        0 5.743590
## 5428      9       9         0                        0 5.090909
## 5429      6      13         0                        0 4.977778
```

```
rm(hflights)
```

```
library(hflights)
hflights %>% head()
```

```
##      Year Month DayofMonth DayOfWeek DepTime ArrTime UniqueCarrier FlightNum
## 5424 2011     1          1         6    1400    1500            AA       428
## 5425 2011     1          2         7    1401    1501            AA       428
## 5426 2011     1          3         1    1352    1502            AA       428
## 5427 2011     1          4         2    1403    1513            AA       428
## 5428 2011     1          5         3    1405    1507            AA       428
## 5429 2011     1          6         4    1359    1503            AA       428
##      TailNum ActualElapsedTime AirTime ArrDelay DepDelay Origin Dest Distance
## 5424  N576AA                60      40      -10        0    IAH  DFW      224
## 5425  N557AA                60      45       -9        1    IAH  DFW      224
## 5426  N541AA                70      48       -8       -8    IAH  DFW      224
## 5427  N403AA                70      39        3        3    IAH  DFW      224
## 5428  N492AA                62      44       -3        5    IAH  DFW      224
## 5429  N262AA                64      45       -7       -1    IAH  DFW      224
##      TaxiIn TaxiOut Cancelled CancellationCode Diverted
## 5424      7      13         0                        0
## 5425      6       9         0                        0
## 5426      5      17         0                        0
## 5427      9      22         0                        0
## 5428      9       9         0                        0
## 5429      6      13         0                        0
```

```
# Ques. Create a table grouped by Unique Carrier and then summarise each group by taking the
mean of ArrDelay.

hflights %>% group_by(UniqueCarrier) %>%
  summarise(mean(ArrDelay,na.rm=T))
```

```
## # A tibble: 15 × 2
##    UniqueCarrier `mean(ArrDelay, na.rm = T)`
##    <chr>                               <dbl>
##  1 AA                                  0.892
##  2 AS                                  3.19
##  3 B6                                  9.86
##  4 CO                                  6.10
##  5 DL                                  6.08
##  6 EV                                  7.26
##  7 F9                                  7.67
##  8 FL                                  1.85
##  9 MQ                                  7.15
## 10 OO                                  8.69
## 11 UA                                 10.5
## 12 US                                 -0.631
## 13 WN                                  7.59
## 14 XE                                  8.19
## 15 YV                                  4.01
```

syntax:summarise_each()

```
# For each unique carrier, calculate the % of flights cancelled or diverted.

hflights %>% group_by(UniqueCarrier) %>%
  summarise_each(funs(mean(.,na.rm=T)),Cancelled,Diverted)
```

```
## Warning: `summarise_each()` was deprecated in dplyr 0.7.0.
## ℹ Please use `across()` instead.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## Warning: `funs()` was deprecated in dplyr 0.8.0.
## ℹ Please use a list of either functions or lambdas:
##
## # Simple named list: list(mean = mean, median = median)
##
## # Auto named with `tibble::lst()`: tibble::lst(mean, median)
##
## # Using lambdas list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## # A tibble: 15 × 3
##    UniqueCarrier Cancelled Diverted
##    <chr>             <dbl>    <dbl>
##  1 AA              0.0185   0.00185
##  2 AS              0        0.00274
##  3 B6              0.0259   0.00576
##  4 CO              0.00678  0.00263
##  5 DL              0.0159   0.00303
##  6 EV              0.0345   0.00318
##  7 F9              0.00716  0
##  8 FL              0.00982  0.00327
##  9 MQ              0.0290   0.00194
## 10 OO              0.0139   0.00349
## 11 UA              0.0164   0.00241
## 12 US              0.0113   0.00147
## 13 WN              0.0155   0.00229
## 14 XE              0.0155   0.00345
## 15 YV              0.0127   0
```

```
# na.rm command is used to remove all NA entries
```

```
# Ques. For each destination calculate the minimum and maximum arr and depdelays.

hflights %>% group_by(Dest) %>% summarise_each(funs(min(.,na.rm=T),max(.,na.rm=T)),DepDelay,A
rrDelay)
```

```
## Warning: `summarise_each()` was deprecated in dplyr 0.7.0.
## ℹ Please use `across()` instead.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## Warning: `funs()` was deprecated in dplyr 0.8.0.
## ℹ Please use a list of either functions or lambdas:
##
## # Simple named list: list(mean = mean, median = median)
##
## # Auto named with `tibble::lst()`: tibble::lst(mean, median)
##
## # Using lambdas list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## # A tibble: 116 × 5
##    Dest  DepDelay_min ArrDelay_min DepDelay_max ArrDelay_max
##    <chr>        <int>        <int>        <int>        <int>
##  1 ABQ            -14          -26          300          290
##  2 AEX            -17          -34          266          257
##  3 AGS             10            4           10            4
##  4 AMA            -13          -28          304          301
##  5 ANC             -6          -21          292          281
##  6 ASE            -33          -31          269          252
##  7 ATL            -17          -41          730          701
##  8 AUS            -15          -24          240          244
##  9 AVL            -10          -23          325          331
## 10 BFL            -13          -56          225          206
## # ℹ 106 more rows
```

n() function counts the no. of rows in a group n_distinct (vector) count the no. of unique items in that vector

```
# Ques. For each day of the year count the total no. of flights and sort in descending order.

hflights %>% group_by(Month,DayofMonth) %>% summarise(flight_count=n()) %>% arrange(desc(flig
ht_count))
```

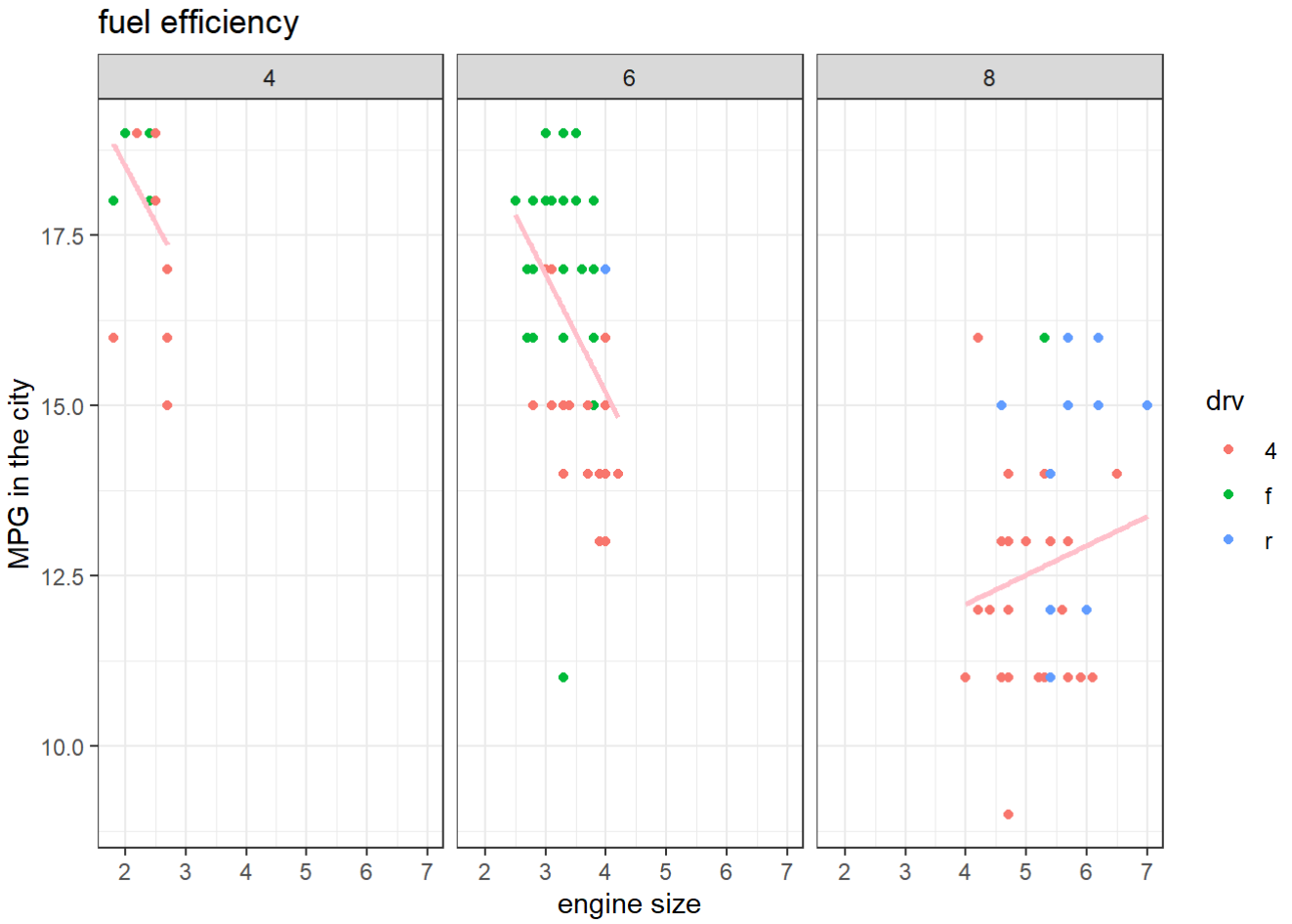
```
## `summarise()` has grouped output by 'Month'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 365 × 3
## # Groups:   Month [12]
##    Month DayofMonth flight_count
##    <int>      <int>        <int>
##  1     8          4          706
##  2     8         11          706
##  3     8         12          706
##  4     8          5          705
##  5     8          3          704
##  6     8         10          704
##  7     1          3          702
##  8     7          7          702
##  9     7         14          702
## 10     7         28          701
## # ℹ 355 more rows
```

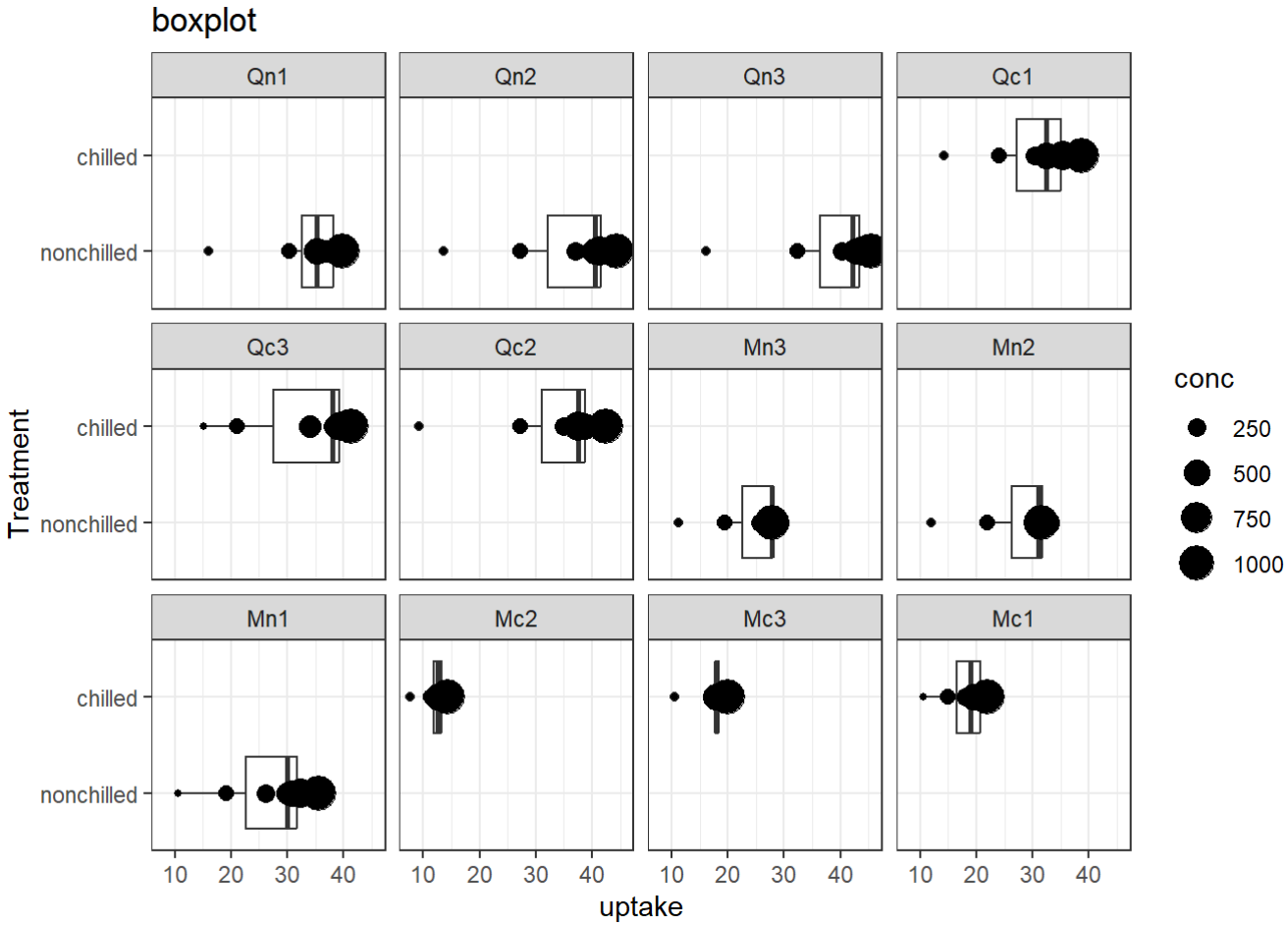
GGPLOT package

```
library(ggplot2)
# example of mpg
# Ques. Plot something for the data where cty<20
library(ggplot2)
mpg %>% filter(cty<20) %>% ggplot(aes(displ,cty))+
geom_point(aes(color=drv))+
geom_smooth(method=lm,se=F,color="pink")+
facet_wrap(~cyl)+
labs(x="engine size",y="MPG in the city",title="fuel efficiency")+
theme_bw()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

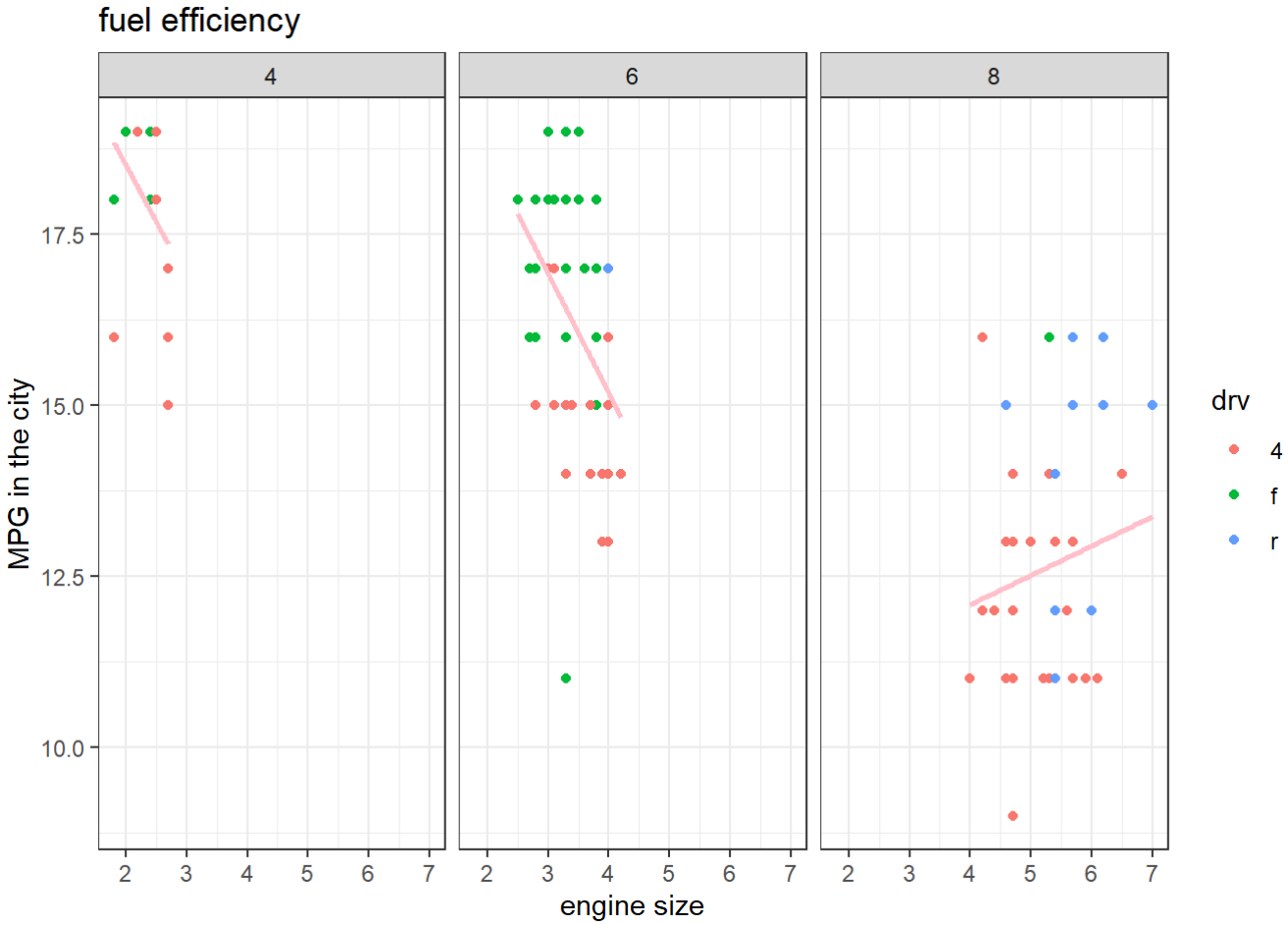



```
# Box Plot
# example of CO2
library(ggplot2)
CO2 %>% ggplot(aes(Treatment,uptake))+
geom_boxplot()+
geom_point(aes(size=conc))+
facet_wrap(~Plant)+
labs(title="boxplot")+
coord_flip()+  # to get horizontal
theme_bw()
```




```
mpg %>% filter(cty<20) %>% ggplot(aes(displ,cty))+
geom_point(aes(color=drv))+
geom_smooth(method=lm,se=F,color="pink")+
facet_wrap(~cyl)+
labs(x="engine size",y="MPG in the city",title="fuel efficiency")+
theme_bw()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```


fuel efficiency
4
6
8
MPG in the city
17.5
15.0
12.5
10.0
2
3
4
5
6
7
engine size
drv
4
f
r