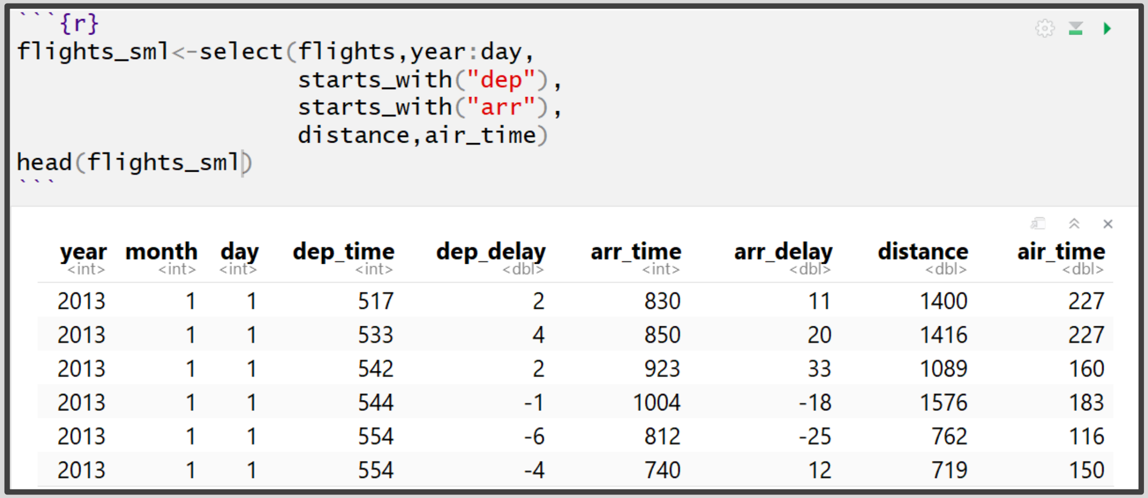
Data Transformation II

mutate()

- Used to Create New Variables
  - Creative New Metrics
  - Modify Units
  - Transform Variables
  - Unique Identifiers
  - Numeric to Categorical
  - Categorical to Numeric

- Reduced Dataset

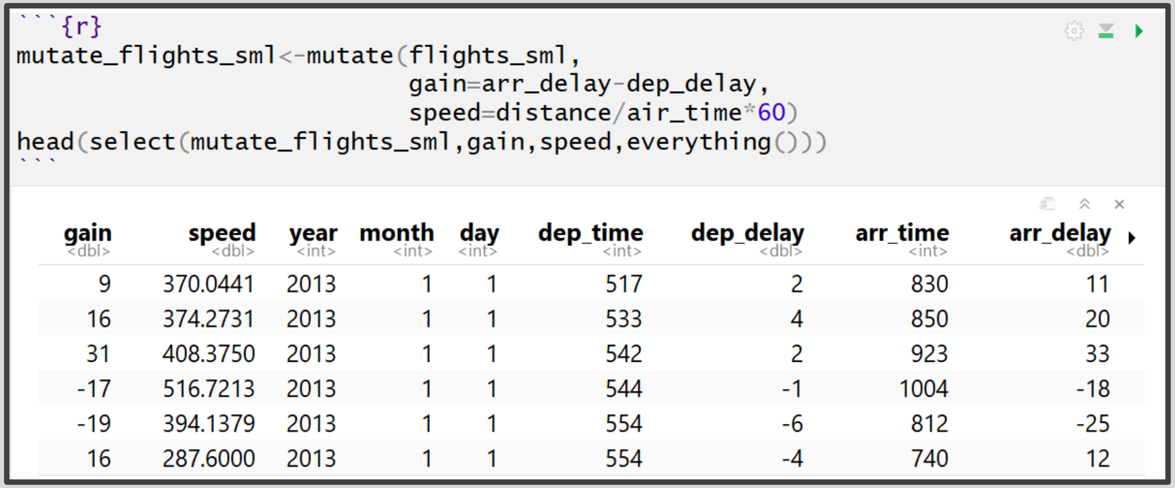


~~~
```{r}
flights_sml<-select(flights,year:day,
                    starts_with("dep"),
                    starts_with("arr"),
                    distance,air_time)
head(flights_sml)
```
~~~

year <int>	month <int>	day <int>	dep_time <int>	dep_delay <dbl>	arr_time <int>	arr_delay <dbl>	distance <dbl>	air_time <dbl>
2013	1	1	517	2	830	11	1400	227
2013	1	1	533	4	850	20	1416	227
2013	1	1	542	2	923	33	1089	160
2013	1	1	544	-1	1004	-18	1576	183
2013	1	1	554	-6	812	-25	762	116
2013	1	1	554	-4	740	12	719	150

mutate()

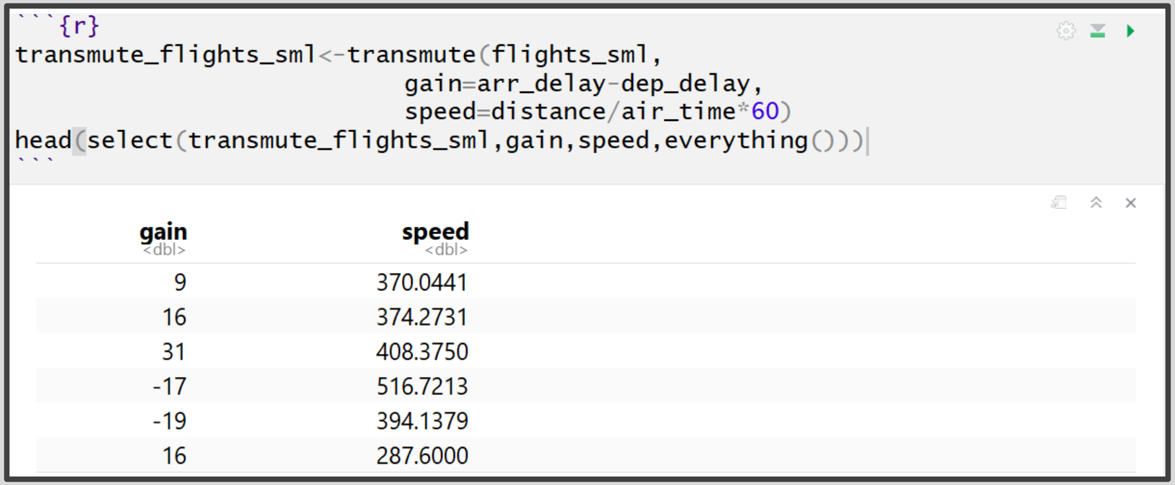
- Example of mutate()



~~~
```{r}
mutate_flights_sml<-mutate(flights_sml,
                          gain=arr_delay-dep_delay,
                          speed=distance/air_time*60)
head(select(mutate_flights_sml,gain,speed,everything()))
```
~~~

gain <dbl>	speed <dbl>	year <int>	month <int>	day <int>	dep_time <int>	dep_delay <dbl>	arr_time <int>	arr_delay <dbl>
9	370.0441	2013	1	1	517	2	830	11
16	374.2731	2013	1	1	533	4	850	20
31	408.3750	2013	1	1	542	2	923	33
-17	516.7213	2013	1	1	544	-1	1004	-18
-19	394.1379	2013	1	1	554	-6	812	-25
16	287.6000	2013	1	1	554	-4	740	12

- Example of transmute()

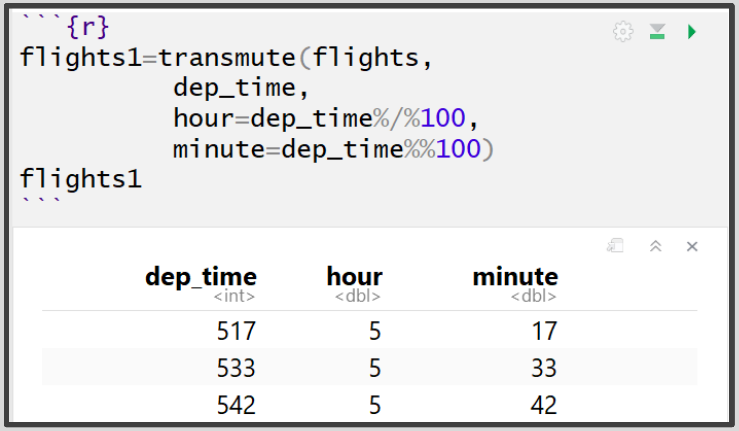


~~~
```{r}
transmute_flights_sml<-transmute(flights_sml,
                         gain=arr_delay-dep_delay,
                         speed=distance/air_time*60)
head(select(transmute_flights_sml,gain,speed,everything()))
```
~~~

gain <dbl>	speed <dbl>
9	370.0441
16	374.2731
31	408.3750
-17	516.7213
-19	394.1379
16	287.6000

mutate()

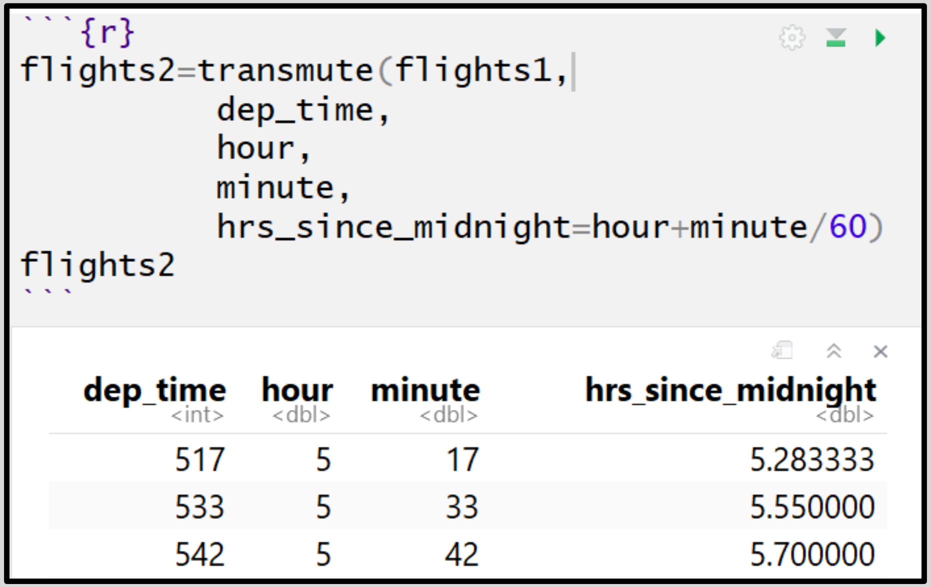
- Plethora of Examples
  - Basic and Modular Arithmetic



~~~
```{r}
flights1=transmute(flights,
          dep_time,
          hour=dep_time%/%100,
          minute=dep_time%%100)
flights1
```
~~~

dep_time <int>	hour <dbl>	minute <dbl>
517	5	17
533	5	33
542	5	42

$$517 = 100 * 5 + 17$$
$$= 100 * (517 \,\%/\%\, 100) + (517 \,\%\%\, 100)$$

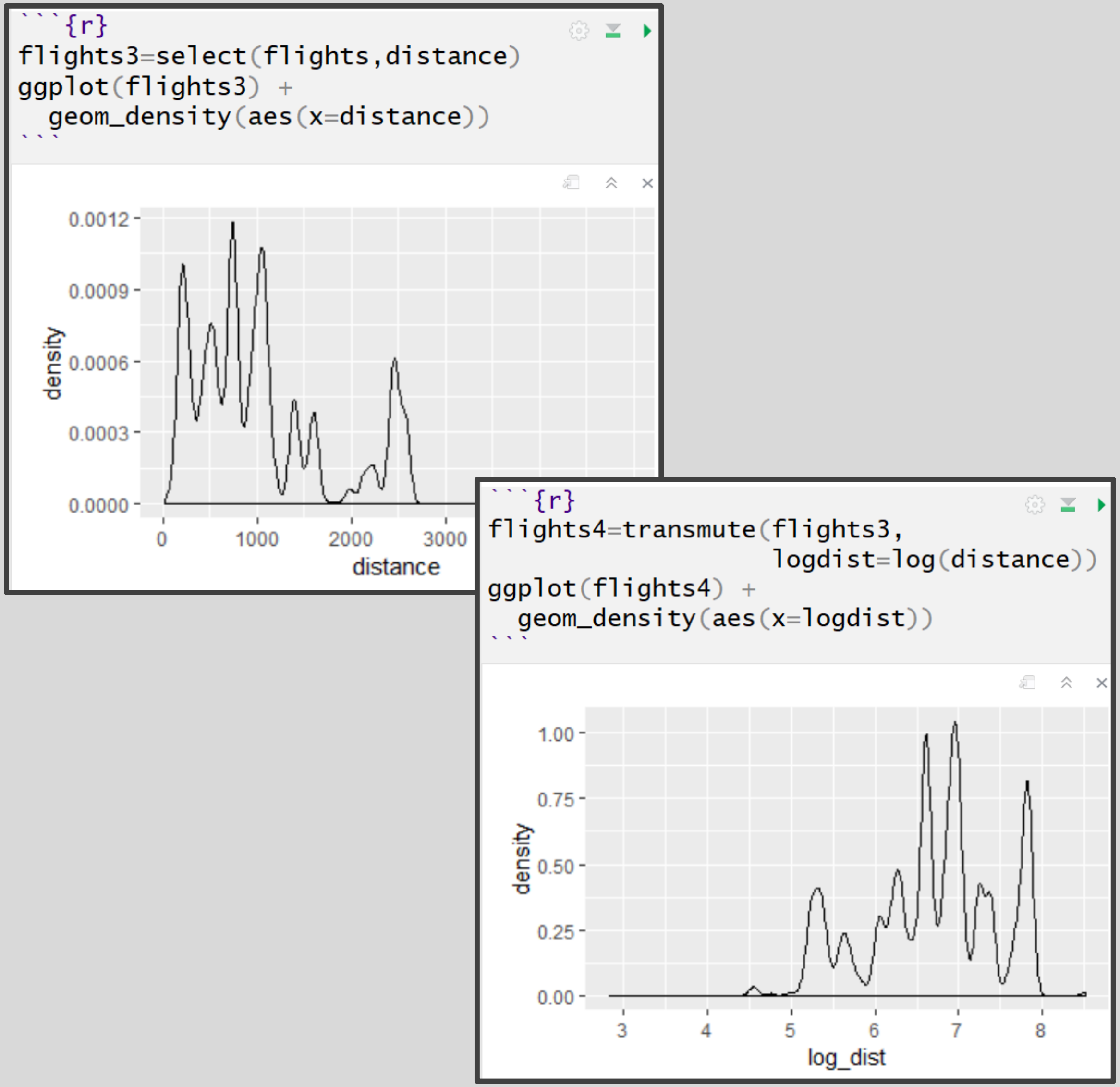


~~~
```{r}
flights2=transmute(flights1,
          dep_time,
          hour,
          minute,
          hrs_since_midnight=hour+minute/60)
flights2
```
~~~

dep_time <int>	hour <dbl>	minute <dbl>	hrs_since_midnight <dbl>
517	5	17	5.283333
533	5	33	5.550000
542	5	42	5.700000

mutate()

- Plethora of Examples
  - Nonlinear Transformation

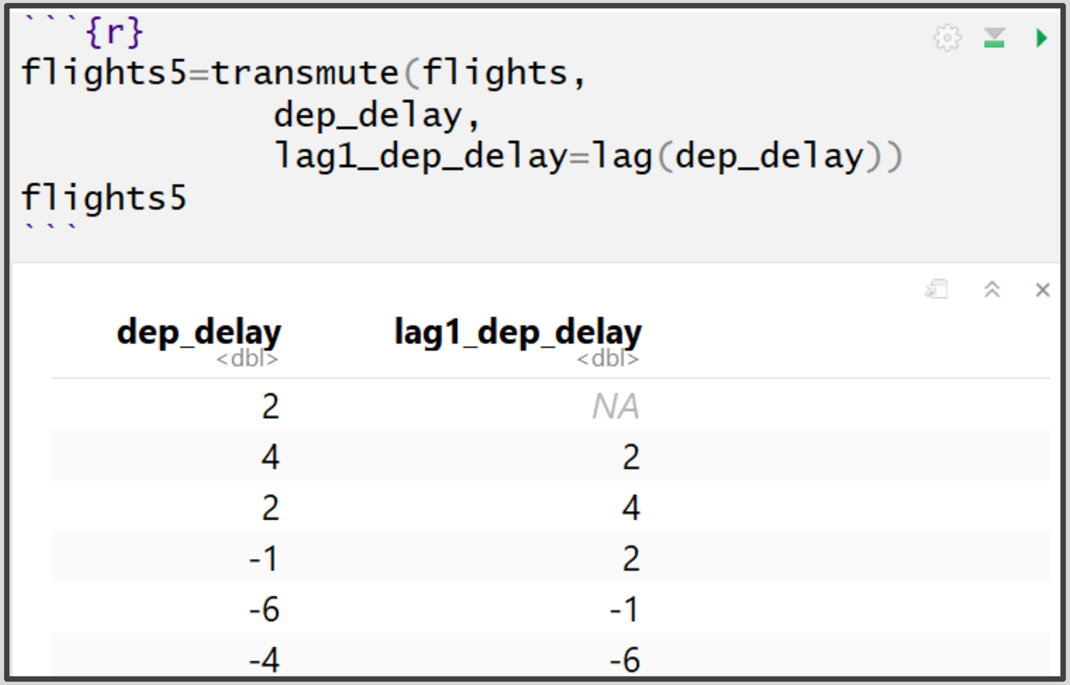


~~~
```{r}
flights3=select(flights,distance)
ggplot(flights3) +
  geom_density(aes(x=distance))
```
~~~

~~~
```{r}
flights4=transmute(flights3,
                   logdist=log(distance))
ggplot(flights4) +
  geom_density(aes(x=logdist))
```
~~~

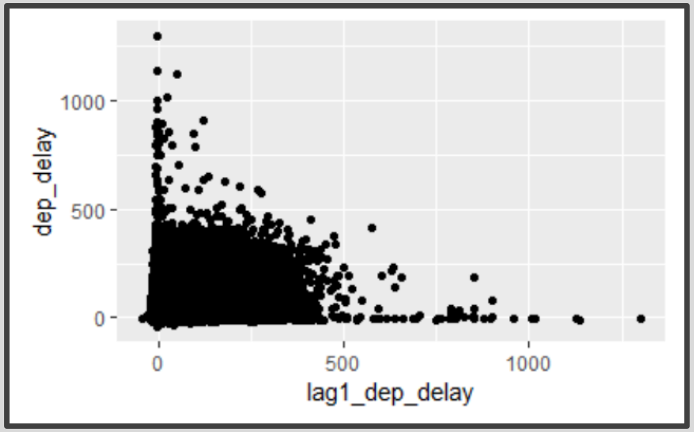
mutate()

- Plethora of Examples
  - Offsets



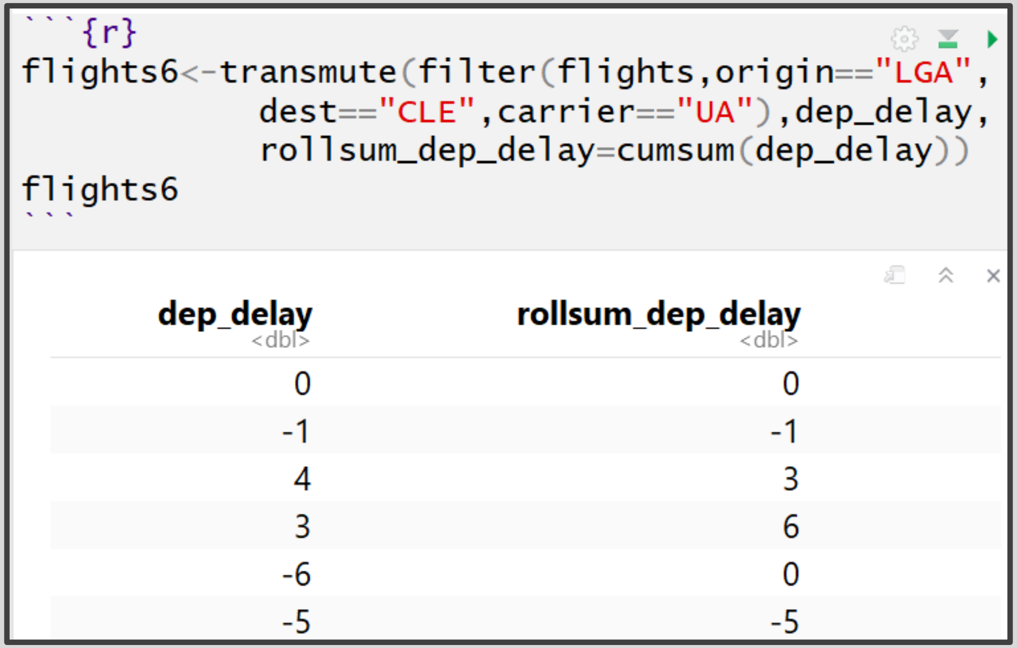
~~~
```{r}
flights5=transmute(flights,
                   dep_delay,
                   lag1_dep_delay=lag(dep_delay))
flights5
```
~~~

```
 dep_delay    lag1_dep_delay
     <dbl>             <dbl>
         2                NA
         4                 2
         2                 4
        -1                 2
        -6                -1
        -4                -6
```



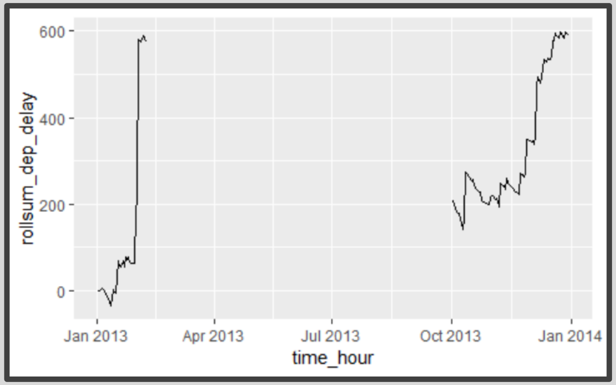
mutate()

- Plethora of Examples
  - Cumulative and Rolling Aggregates



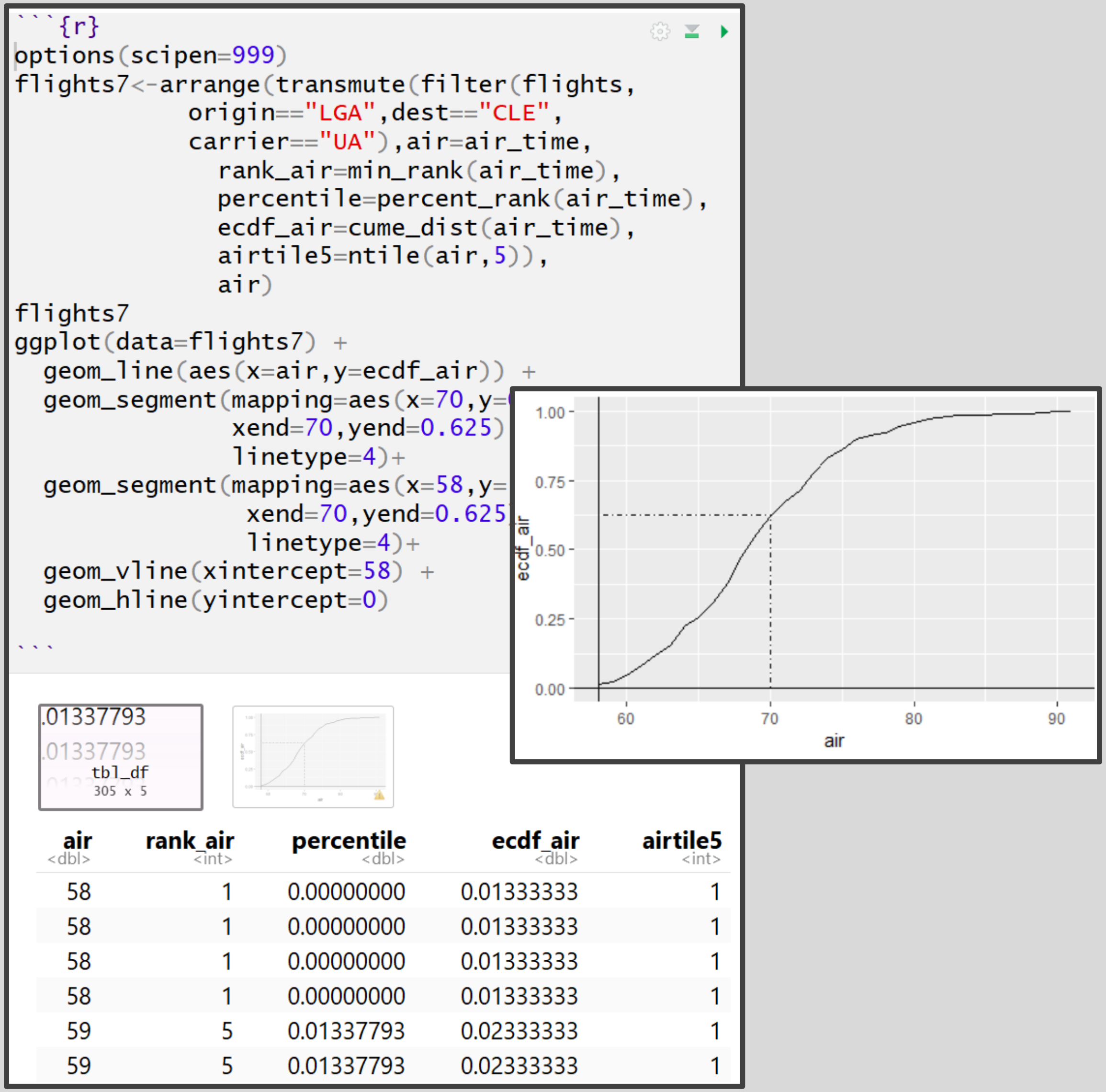
~~~
```{r}
flights6<-transmute(filter(flights,origin=="LGA",
           dest=="CLE",carrier=="UA"),dep_delay,
           rollsum_dep_delay=cumsum(dep_delay))
flights6
```
~~~

dep_delay <dbl>	rollsum_dep_delay <dbl>
0	0
-1	-1
4	3
3	6
-6	0
-5	-5



mutate()

- Plethora of Examples
  - Ranking



~~~
```{r}
options(scipen=999)
flights7<-arrange(transmute(filter(flights,
            origin=="LGA",dest=="CLE",
            carrier=="UA"),air=air_time,
              rank_air=min_rank(air_time),
              percentile=percent_rank(air_time),
              ecdf_air=cume_dist(air_time),
              airtile5=ntile(air,5)),
              air)
flights7
ggplot(data=flights7) +
  geom_line(aes(x=air,y=ecdf_air)) +
  geom_segment(mapping=aes(x=70,y=
               xend=70,yend=0.625)
               linetype=4)+
  geom_segment(mapping=aes(x=58,y=
                xend=70,yend=0.625
                linetype=4)+
  geom_vline(xintercept=58) +
  geom_hline(yintercept=0)

```
~~~

.01337793
.01337793
tbl_df
305 x 5

air <dbl>	rank_air <int>	percentile <dbl>	ecdf_air <dbl>	airtile5 <int>
58	1	0.00000000	0.01333333	1
58	1	0.00000000	0.01333333	1
58	1	0.00000000	0.01333333	1
58	1	0.00000000	0.01333333	1
59	5	0.01337793	0.02333333	1
59	5	0.01337793	0.02333333	1

Information

- Tutorial 3
  - Practice
    - filter()
    - arrange()
    - select()
    - mutate()

  - Introduced
    - Piping %>%
    - group_by()
    - summarize()

- Chain of Command

Google -> Friend -> Instructor

Closing

Disperse and Make Reasonable Decisions