Data Transformation IV

Case Study

- Flight Accuracy
  - Accurate Flight Means
    - Departure Delay = 0
    - Arrival Delay = 0
  - Bad Metric

$$Accuracy = delay_{dep} + delay_{arr}$$
$$Accuracy = (delay_{dep} + delay_{arr})/2$$

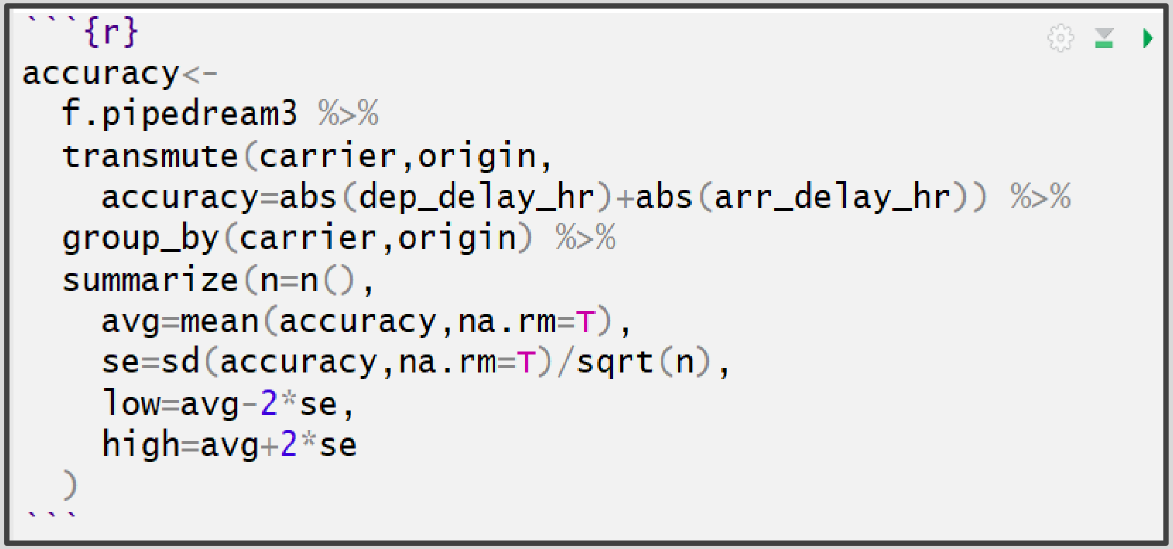
  - Good Metrics

$$Accuracy = |delay_{dep}| + |delay_{arr}|$$
$$Accuracy = \sqrt{delay_{dep}^2 + delay_{arr}^2}$$

  - Table First, Graphics Second

Case Study

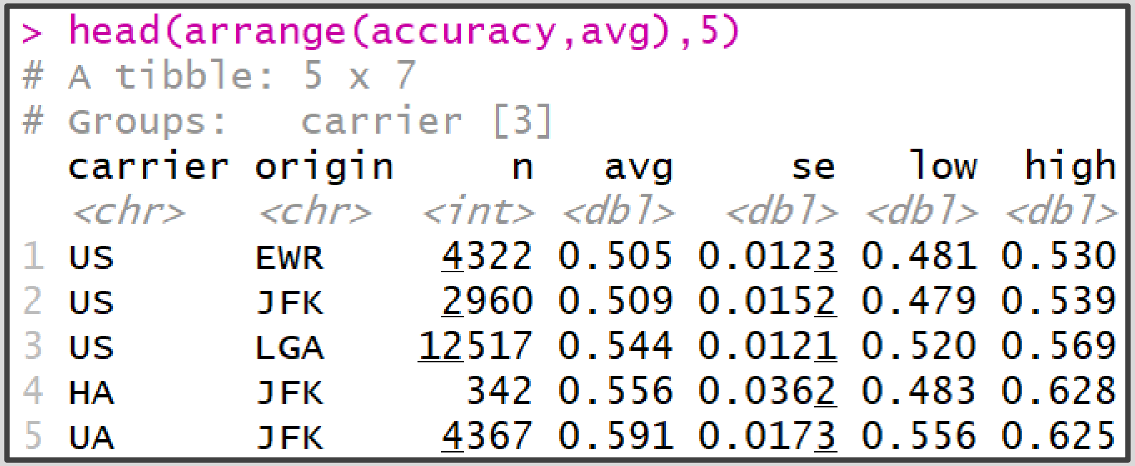
- Summary Table
  - Step 1: Accuracy Variable
  - Step 2: Grouping
  - Step 3: Summarize Info
    - Mean
    - Standard Error
    - Lower Bound (95% CI)
    - Upper Bound (95% CI)



~~~
```{r}
accuracy<-
  f.pipedream3 %>%
  transmute(carrier,origin,
    accuracy=abs(dep_delay_hr)+abs(arr_delay_hr)) %>%
  group_by(carrier,origin) %>%
  summarize(n=n(),
    avg=mean(accuracy,na.rm=T),
    se=sd(accuracy,na.rm=T)/sqrt(n),
    low=avg-2*se,
    high=avg+2*se
  )
```
~~~

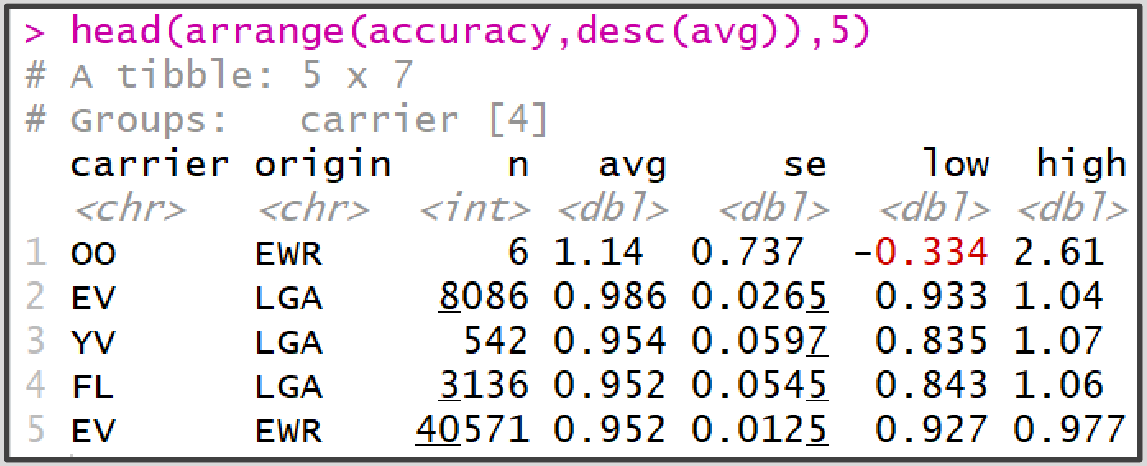
Case Study

- Sorted by Average Accuracy
  - Best Carriers/Origin



```
> head(arrange(accuracy,avg),5)
# A tibble: 5 x 7
# Groups:   carrier [3]
  carrier origin     n   avg     se   low  high
  <chr>   <chr>  <int> <dbl>  <dbl> <dbl> <dbl>
1 US      EWR     4322 0.505 0.0123 0.481 0.530
2 US      JFK     2960 0.509 0.0152 0.479 0.539
3 US      LGA    12517 0.544 0.0121 0.520 0.569
4 HA      JFK      342 0.556 0.0362 0.483 0.628
5 UA      JFK     4367 0.591 0.0173 0.556 0.625
```

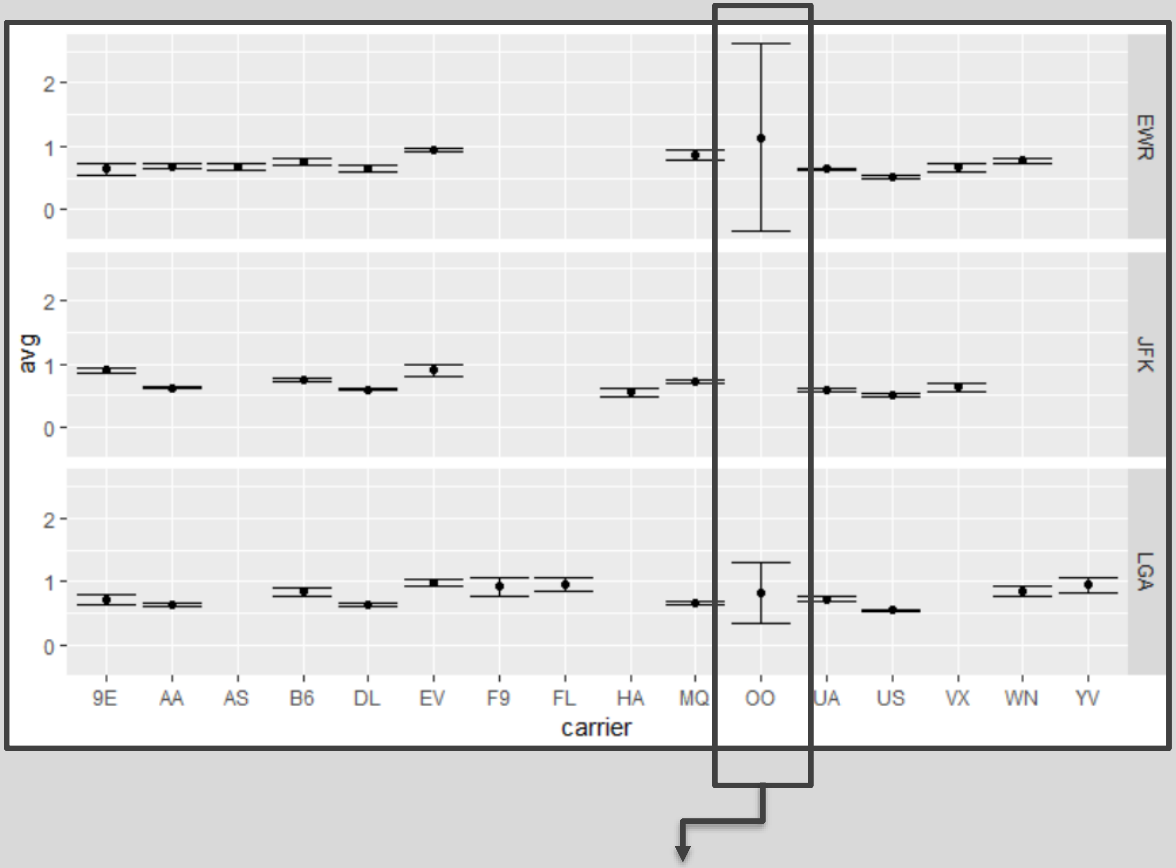
- Worst Carriers/Origin



```
> head(arrange(accuracy,desc(avg)),5)
# A tibble: 5 x 7
# Groups:   carrier [4]
  carrier origin     n   avg     se    low  high
  <chr>   <chr>  <int> <dbl>  <dbl>  <dbl> <dbl>
1 OO      EWR        6 1.14  0.737  -0.334 2.61
2 EV      LGA     8086 0.986 0.0265  0.933 1.04
3 YV      LGA      542 0.954 0.0597  0.835 1.07
4 FL      LGA     3136 0.952 0.0545  0.843 1.06
5 EV      EWR    40571 0.952 0.0125  0.927 0.977
```

Case Study

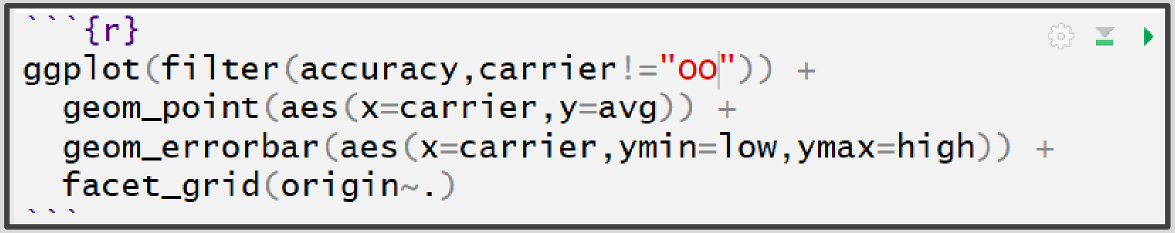
- 95% Confidence Intervals



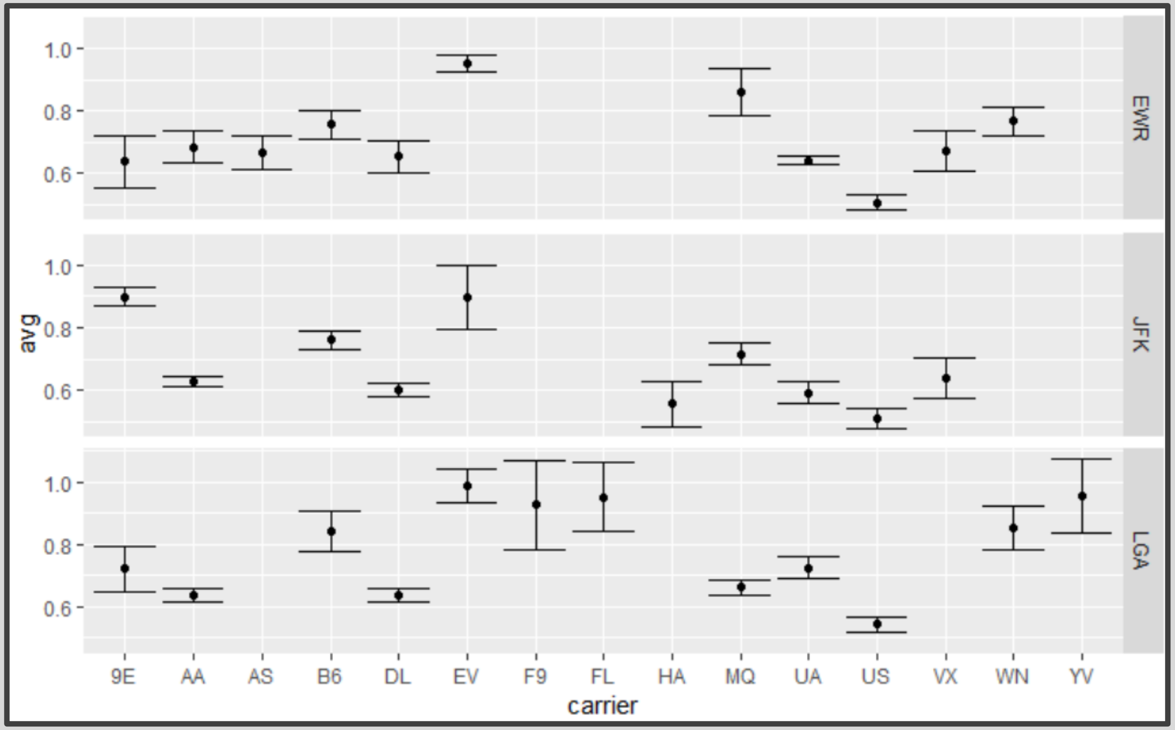
Carrier “OO” Creates a Visual Problem Due to Small Sample Size

Case Study

- 95% Confidence Intervals



~~~
```{r}
ggplot(filter(accuracy,carrier!="OO")) +
  geom_point(aes(x=carrier,y=avg)) +
  geom_errorbar(aes(x=carrier,ymin=low,ymax=high)) +
  facet_grid(origin~.)
```
~~~



Closing

Disperse and Make Reasonable Decisions