Midterm 2 Prep - Code Along

Wendy Moore

2024-11-05

Code	Airline
9E	Endeavor Air Inc.
AA	American Airlines Inc.
AS	Alaska Airlines Inc.
B6	JetBlue Airways
DL	Delta Air Lines Inc.
EV	ExpressJet Airlines Inc.
F9	Frontier Airlines Inc.
FL	AirTran Airways Corporation
HA	Hawaiian Airlines Inc.
MQ	Envoy Air
OO	SkyWest Airlines Inc.
UA	United Air Lines Inc.
US	US Airways Inc.
VX	Virgin America
WN	Southwest Airlines Co.
YV	Mesa Airlines Inc.

Missing values

```
knitr::kable(apply(nycflights, 2, function(x) sum(is.na(x))))
```

	x
year	0
month	0
day	0
dep_time	0
dep_delay	0
arr_time	0
arr_delay	0
carrier	0
tailnum	0
flight	0
origin	0
dest	0
air_time	0
distance	0
hour	0
minute	0

Exclusion Criteria

```
table(nycflights$carrier)
```

```
##
##   9E   AA   AS   B6   DL   EV   F9   FL   HA   MQ   OO   UA   US   VX   WN   YV
## 1696 3188   66 5376 4751 5142   69  307   34 2507    3 5770 2015  497 1261   53
```

```
nycflights <- nycflights %>%
  filter(carrier != "OO")

nycflights <- nycflights %>%
  mutate_if(is.character, as.factor)
```

EDA - Descriptive Statistics

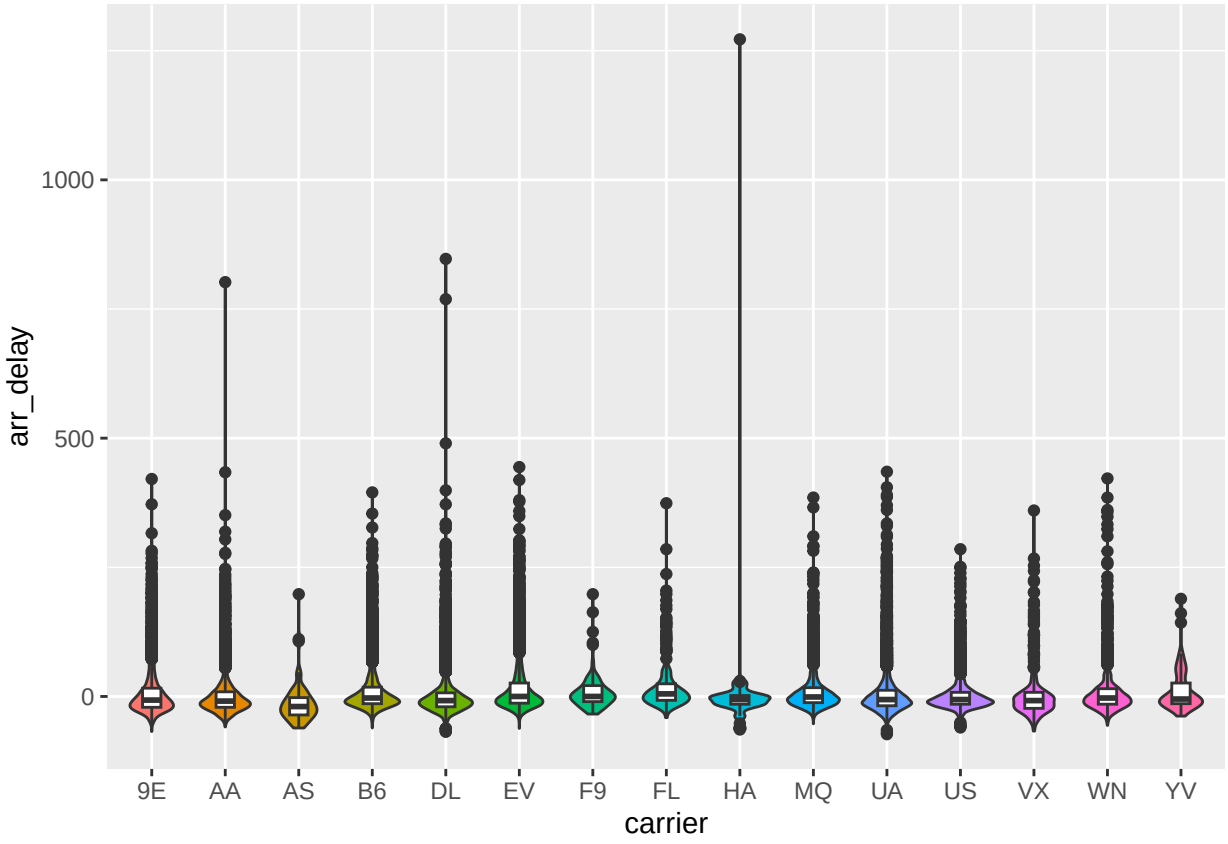
```
nycflights %>%
  select(origin, arr_delay, carrier, distance, air_time) %>%
  tbl_summary(by = origin,
              statistic = list(all_continuous() ~ "{mean} ({sd})"),
              type = list(all_dichotomous() ~ "categorical"),
              digits = list(all_continuous() ~ c(2,2)),
              label = list(
                arr_delay = "Arrival Delay (in min)",
                carrier = "Carrier",
                distance = "Distance",
                air_time = "Air Time (min)")) %>%
  modify_header(stat_1 = "Newark",
                stat_2 = "JFK",
                stat_3 = "Laguardia")
```

EDA - Data Visualization

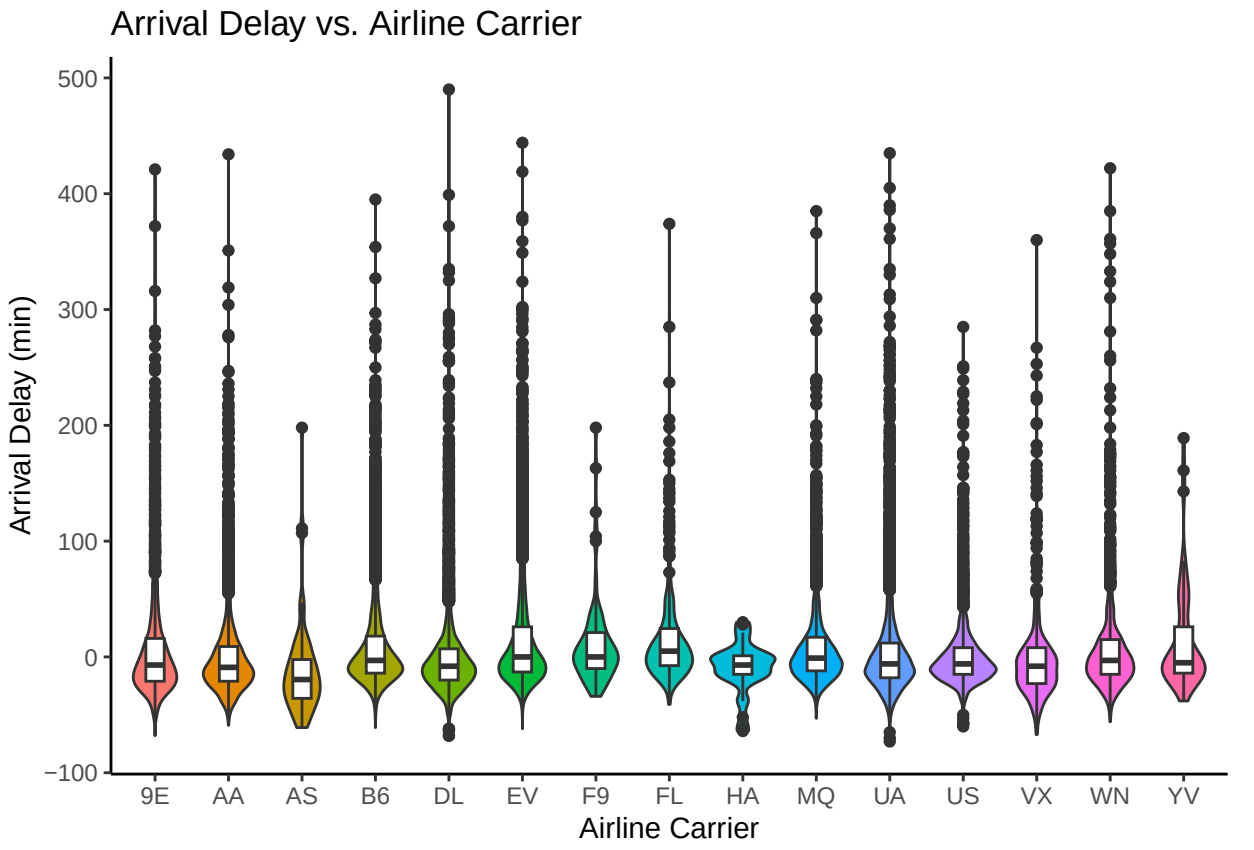
```
ggplot(nycflights, aes(x = carrier, y = arr_delay)) +
  geom_violin(aes(fill = carrier),
              show.legend = FALSE) +
  geom_boxplot(width = 0.25)
```

Characteristic	Newark[1]	JFK[1]	Laguardia[1]
Arrival Delay (in min)	9.33 (46.09)	5.98 (44.52)	5.70 (43.09)
Carrier			
9E	121 (1.0%)	1,314 (12%)	261 (2.6%)
AA	350 (3.0%)	1,388 (13%)	1,450 (14%)
AS	66 (0.6%)	0 (0%)	0 (0%)
B6	625 (5.3%)	4,166 (38%)	585 (5.8%)
DL	445 (3.8%)	2,070 (19%)	2,236 (22%)
EV	4,170 (35%)	118 (1.1%)	854 (8.5%)
F9	0 (0%)	0 (0%)	69 (0.7%)
FL	0 (0%)	0 (0%)	307 (3.1%)
HA	0 (0%)	34 (0.3%)	0 (0%)
MQ	210 (1.8%)	717 (6.6%)	1,580 (16%)
UA	4,559 (39%)	440 (4.0%)	771 (7.7%)
US	444 (3.8%)	302 (2.8%)	1,269 (13%)
VX	149 (1.3%)	348 (3.2%)	0 (0%)
WN	631 (5.4%)	0 (0%)	630 (6.3%)
YV	0 (0%)	0 (0%)	53 (0.5%)
Distance	1,058.45 (726.52)	1,273.92 (895.70)	785.60 (375.05)
Air Time (min)	152.44 (92.58)	178.27 (113.90)	117.99 (49.83)

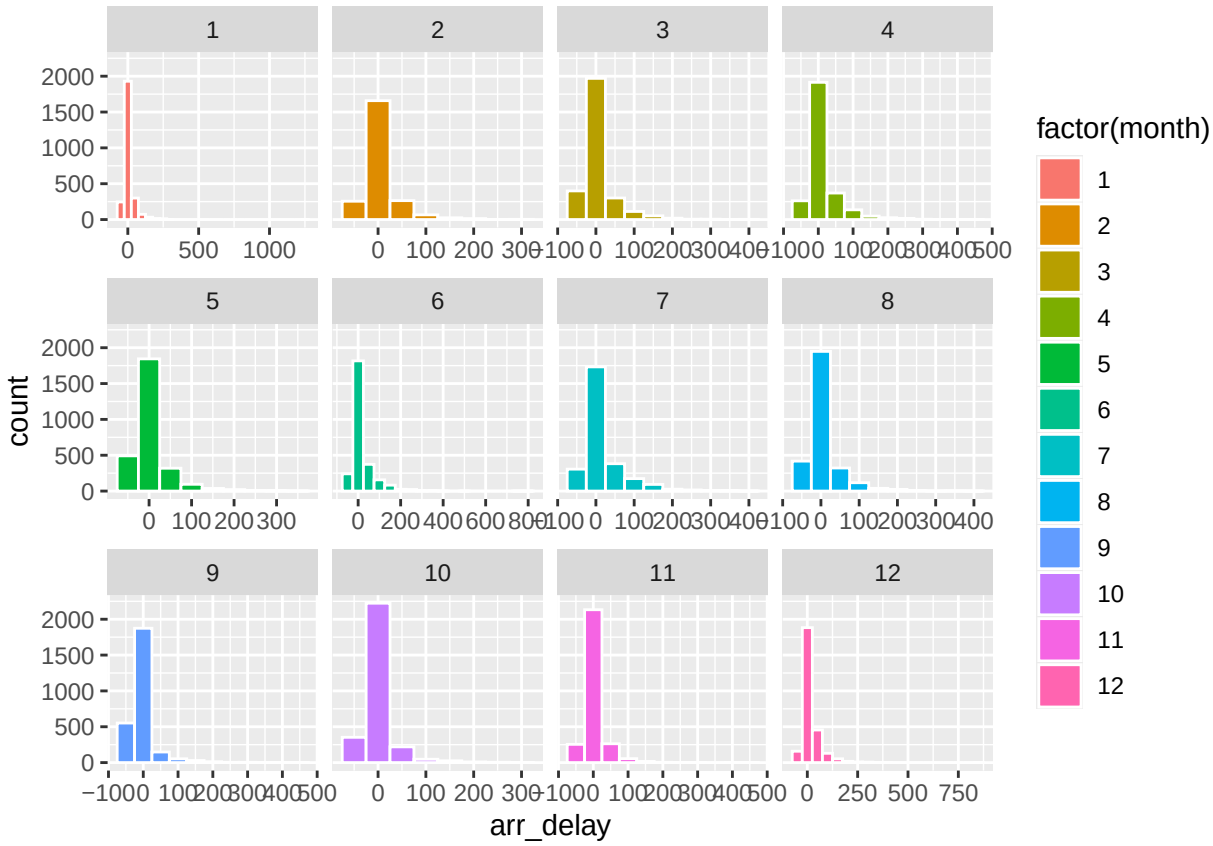
[1]Mean (SD); n (%)



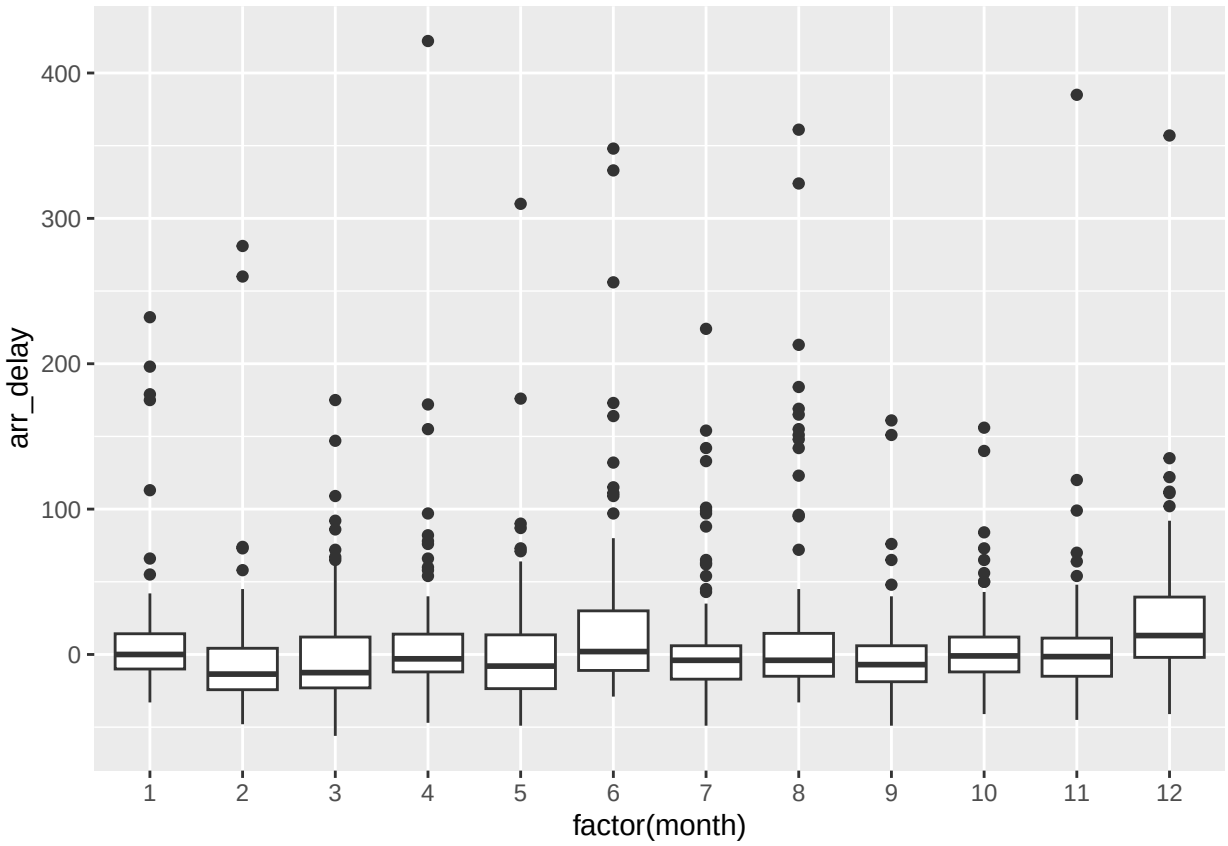
```r
nycflights %>%
  filter(arr_delay < 500) %>%
  ggplot(aes(x = carrier, y = arr_delay)) +
  geom_violin(aes(fill = carrier),
              show.legend = FALSE) +
  geom_boxplot(width = 0.25) +
  labs(title = "Arrival Delay vs. Airline Carrier",
       x = "Airline Carrier",
       y = "Arrival Delay (min)") +
  theme_classic()
```




```r
ggplot(nycflights, aes(x = arr_delay, fill = factor(month))) +
  geom_histogram(binwidth = 50, col = "white") +
  facet_wrap(~ month, scales = "free_x")
```




```
nycflights %>%
  filter(carrier == "WN") %>%
  ggplot(aes(x = factor(month), y = arr_delay)) +
  geom_boxplot()
```



One Proportion

Is the proportion of Southwest Airlines (WN) flights that depart late (>15 min) less than 50%?

```
nycflights %>%
  mutate(late_dep = ifelse(dep_delay > 15, 1, 0)) %>%
  filter(carrier == "WN") %>%
  group_by(late_dep) %>%
  summarise(n = n())
```

```
## # A tibble: 2 x 2
##   late_dep     n
##      <dbl> <int>
## 1        0   925
## 2        1   336
```

```
one_prop_test <- prop.test(336, (925 + 336), p = 0.5, alternative = "less",
          correct = FALSE)

# Z-statistic
-sqrt(one_prop_test$statistic)
```

```
## X-squared
## -16.58661
```

```
# P-value
one_prop_test$p.value
```

```
## [1] 4.354668e-62
```

```r
### OR by hand
n <- 1261
phat <- 336 / n
p0 <- 0.5

# Check conditions (both counts >= 10)
n*p0
```

```
## [1] 630.5
```

```r
n*(1-p0)
```

```
## [1] 630.5
```

```r
se_null <- sqrt( (p0 * (1 - p0) ) / n)

z_stat <- (phat - p0) / se_null
z_stat
```

```
## [1] -16.58661
```

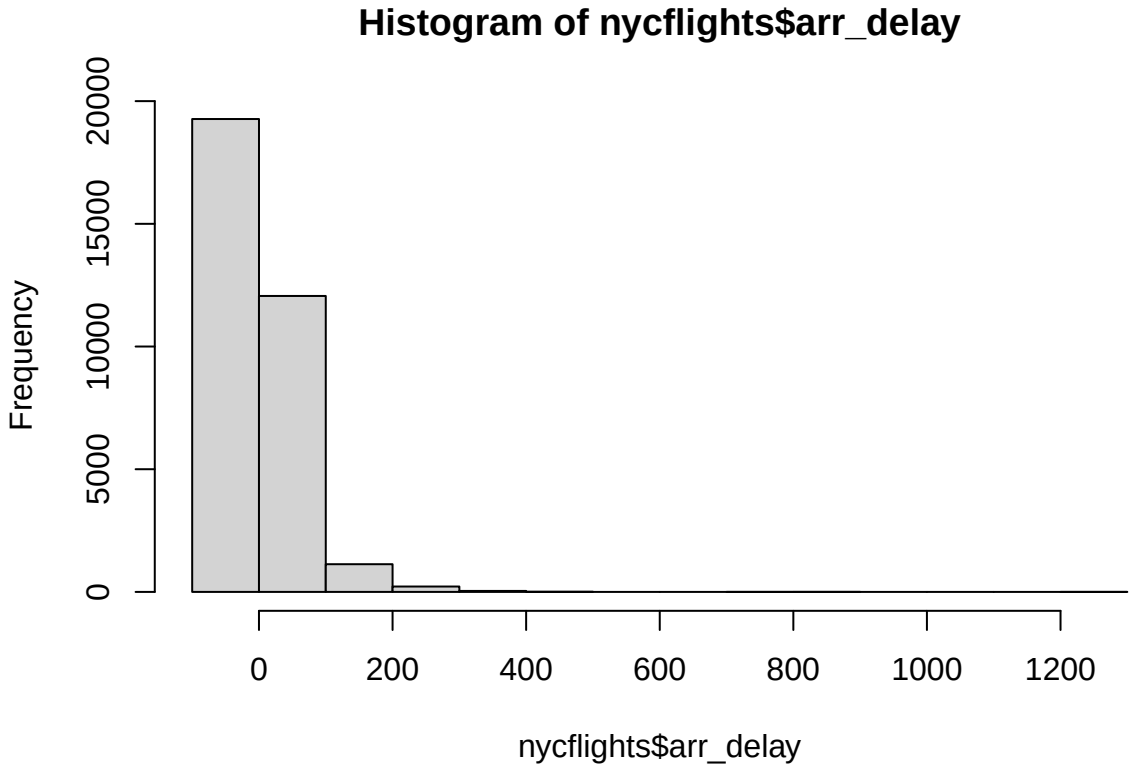
```r
p_value <- pnorm(z_stat)
p_value
```

```
## [1] 4.354668e-62
```

One Mean

Across all airlines, is the average arrival minutes late different from 0 (i.e., on time)?

```r
# Yikes... super skewed!
hist(nycflights$arr_delay)
```



```r
t.test(nycflights$arr_delay)
```

```
##
##  One Sample t-test
##
## data:  nycflights$arr_delay
## t = 28.744, df = 32731, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  6.615797 7.584069
## sample estimates:
## mean of x
##  7.099933
```

Difference in Proportions

Are the differences in proportions of late (> 15 min) flight departures difference for Southwest compared to Delta Airlines?

```r
nycflights %>%
  filter(carrier == "WN" | carrier == "DL") %>%
  mutate(late_dep = ifelse(dep_delay > 15, 1, 0)) %>%
  group_by(carrier, late_dep) %>%
  summarise(n())
```

```
## # A tibble: 4 x 3
## # Groups:   carrier [2]
##   carrier late_dep `n()`
##   <fct>      <dbl> <int>
## 1 DL             0  4044
## 2 DL             1   707
## 3 WN             0   925
## 4 WN             1   336
```

```r
# Uses pooled proportion
diff_prop <- prop.test(x = c(336, 707), n = c(1261, 4751),
                       correct = FALSE)
sqrt(diff_prop$statistic)
```

```
## X-squared
##  9.807407
```

```r
n1 <- 1261
phat1 <- 336/n1

n2 <- 4751
phat2 <- 707/n2

n1*phat1; n1*(1-phat1)
```

```
## [1] 336
```

```
## [1] 925
```

```r
n2*phat2; n2*(1-phat2)
```

```
## [1] 707
```

```
## [1] 4044
```

```
# Pooled
phat_pooled <- (phat1*n1 + phat2*n2)/ (n1 + n2)

n1*phat_pooled; n1*(1-phat_pooled)
```

```
## [1] 218.7663
```

```
## [1] 1042.234
```

```
n2*phat_pooled; n2*(1-phat_pooled)
```

```
## [1] 824.2337
```

```
## [1] 3926.766
```

```
est <- phat1 - phat2

# Not pooled
se_phats <- sqrt((phat1*(1-phat1))/n1 + (phat2*(1-phat2))/n2)
se_phats
```

```
## [1] 0.01347822
```

```
# Pooled
se_pooled <- sqrt( (phat_pooled * (1 - phat_pooled)) * (1/n1 + 1/n2) )
se_pooled
```

```
## [1] 0.01199547
```

```
# Test statistic not pooled
z_stat <- (est - 0) / se_phats
z_stat
```

```
## [1] 8.728484
```

```
# Test statistic pooled
z_pooled <- (est - 0) / se_pooled
z_pooled
```

```
## [1] 9.807407
```

```
pval <- pnorm(z_stat)
pval
```

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

```
pval_pooled <- pnorm(z_pooled)
pval_pooled
```

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

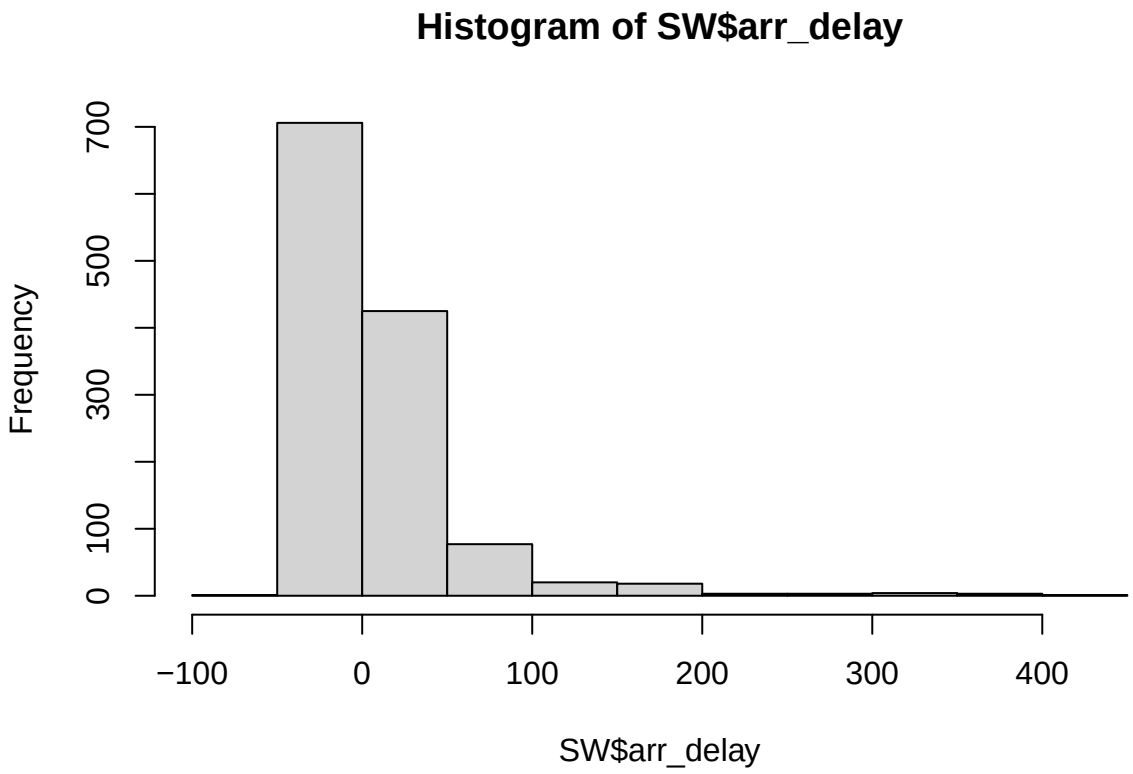
Difference in Means (independent)

Is there a difference in the average minutes arriving late for Southwest compared to Delta Airlines?

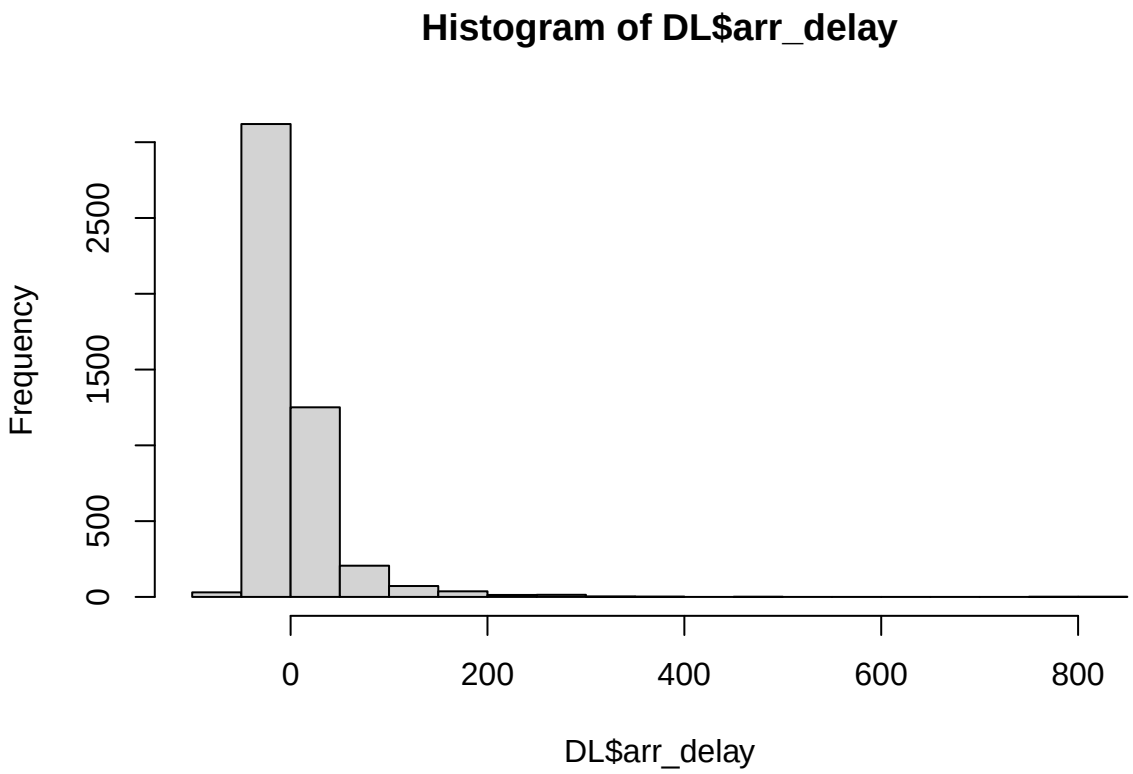
```
SW <- nycflights %>%
  filter(carrier == "WN")

DL <- nycflights %>%
  filter(carrier == "DL")
```

```r
# Skewed but very large sample size
hist(SW$arr_delay)
```




```r
# Very skewed.... results might not be valid
hist(DL$arr_delay)
```




```
# Do the test anyway
t.test(SW$arr_delay, DL$arr_delay)
```

```
##
##  Welch Two Sample t-test
##
## data:  SW$arr_delay and DL$arr_delay
## t = 5.3448, df = 1838.5, p-value = 1.018e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   5.049645 10.903694
## sample estimates:
## mean of x mean of y
## 8.8834259 0.9067565
```

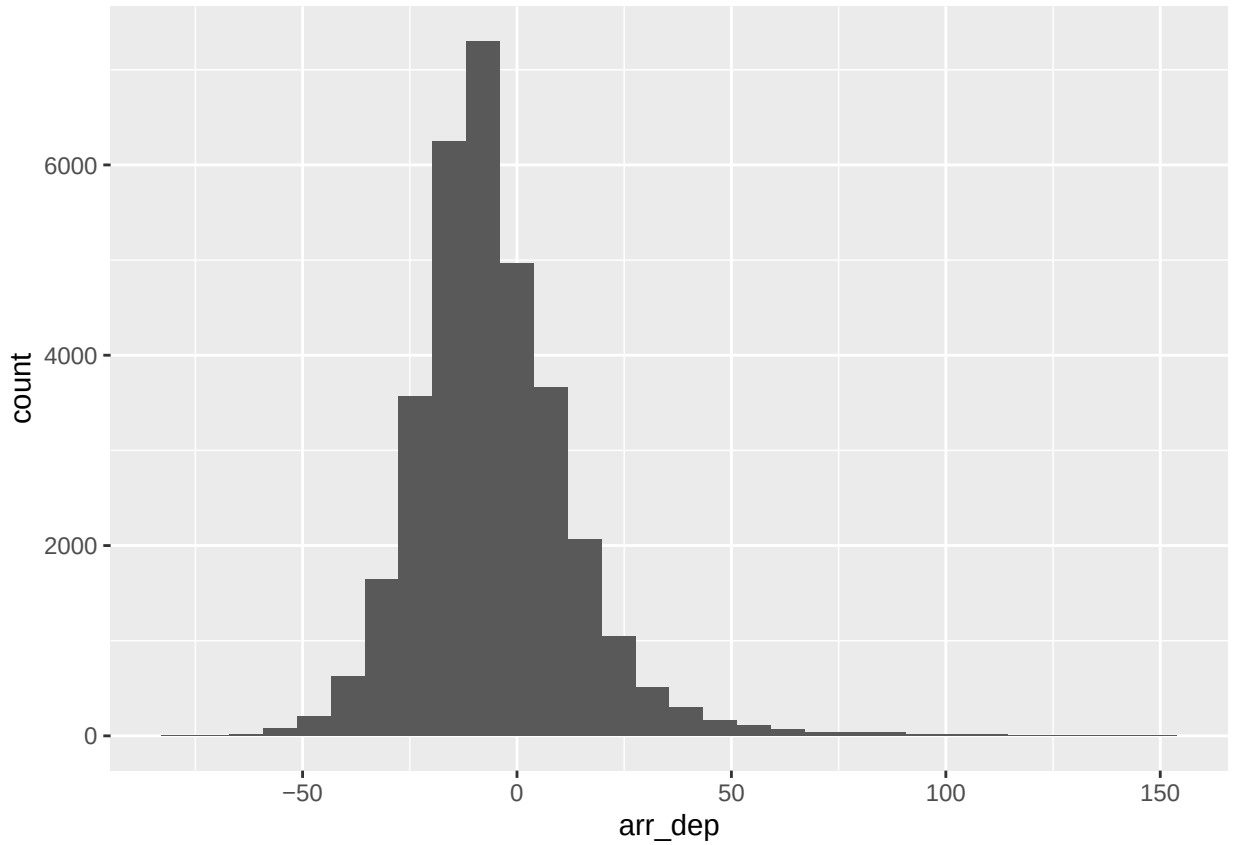
Difference in Means (paired)

$H_0 : \mu_d = 0$

$H_A : \mu_d \neq 0$

Check conditions: 1. Independence: NOT satisfied
2. Normality:

```
nycflights$arr_dep <- nycflights$arr_delay - nycflights$dep_delay
ggplot(nycflights, aes(x = arr_dep)) +
  geom_histogram()
```




```
t.test(nycflights$arr_delay, nycflights$dep_delay,
       paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  nycflights$arr_delay and nycflights$dep_delay
## t = -56.551, df = 32731, p-value < 2.2e-16
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -5.798608 -5.410117
## sample estimates:
## mean difference
##       -5.604363
```

```
## OR
t.test(nycflights$arr_dep)
```

```
##
##  One Sample t-test
##
## data:  nycflights$arr_dep
## t = -56.551, df = 32731, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -5.798608 -5.410117
## sample estimates:
## mean of x
```

```
## -5.604363
```

Chi-squared Test for GoF

Are the proportions of flights leaving each of the 3 NYC airports the same?

```
gof_test <- chisq.test(table(nycflights$origin))
gof_test$expected
```

```
##      EWR      JFK      LGA
## 10910.67 10910.67 10910.67
```

```
gof_test
```

```
##
##  Chi-squared test for given probabilities
##
## data:  table(nycflights$origin)
## X-squared = 133.25, df = 2, p-value < 2.2e-16
```

Chi-squared Test for Independence

H_0: dep delay (categorized) is independent from origin airport
H_A: dep delay (categorized) is NOT independent from origin airport

```
nycflights <- nycflights %>%
  mutate(dep_delay_cat = factor(case_when(dep_delay < 0 ~ "Early",
                                          dep_delay < 15 ~ "On Time",
                                          TRUE ~ "Late"))) %>%
  mutate(dep_delay_cat = forcats::fct_relevel(dep_delay_cat,
                                              c("Early", "On Time", "Late")))

tab <- table(nycflights$origin, nycflights$dep_delay_cat)
tab
```

```
##
##       Early On Time Late
##   EWR  5892    2858 3020
##   JFK  6122    2458 2317
##   LGA  6293    1805 1967
```

```
chisq.test(tab)$expected
```

```
##
##          Early  On Time     Late
##   EWR 6582.958 2560.619 2626.423
##   JFK 6094.690 2370.693 2431.617
##   LGA 5629.352 2189.688 2245.960
```

```
chisq.test(tab)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab
## X-squared = 355.25, df = 4, p-value < 2.2e-16
```