Project1

Fidan Aydamirova

10/23/2021

Using Statistical Hypothesis Testing to find the pattern between departure delay and season/time of a day.

1. Is there a pattern regarding departure delays versus time of day? Fidan

```
library(ggplot2)
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✔ tibble  3.1.8      ✔ dplyr   1.0.10
## ✔ tidyr   1.2.1      ✔ stringr 1.4.1
## ✔ readr   2.1.2      ✔ forcats 0.5.2
## ✔ purrr   0.3.4
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

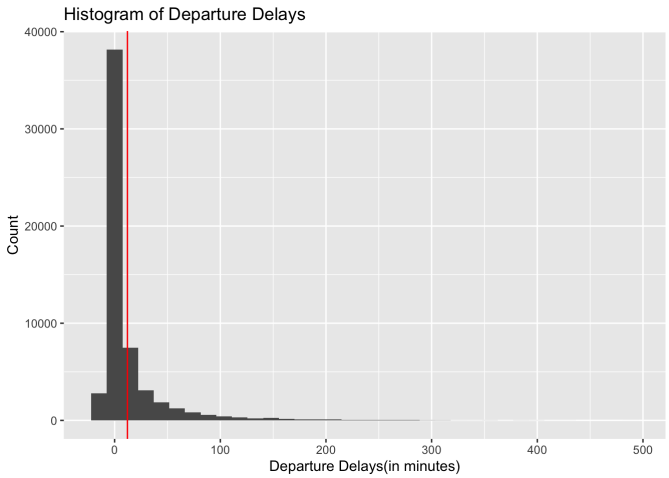
```
library(nycflights13)
```

```
flightsUA <- flights %>%
  filter(carrier == "UA")
```

```
flightsUA <- flightsUA %>% drop_na(dep_delay)
flightsUA <- flightsUA %>% drop_na(sched_dep_time)
```

```
flightsUA<-mutate(flightsUA, timeOfDay = ifelse(sched_dep_time %in% 600:1159, "morning",
                            ifelse(sched_dep_time %in% 1200:1759, "afternoon",
                            ifelse(sched_dep_time %in% 1800:2359, "evening","night"))))
```

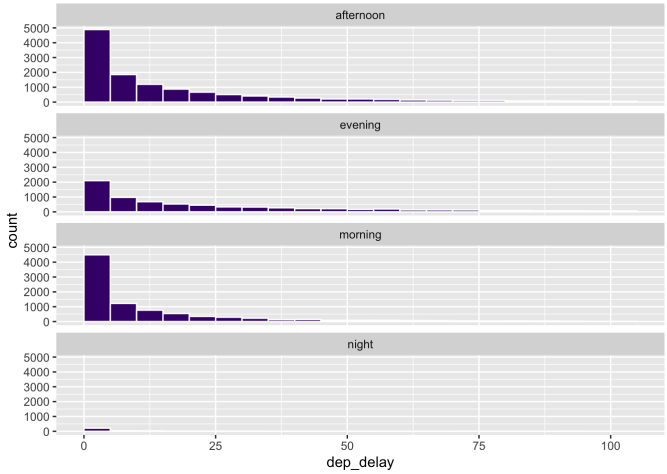
```
#departure delay  distribution of flights by histogram
ggplot(data=flightsUA, mapping = aes(x=dep_delay)) +
geom_histogram(bins = 35)+
geom_vline(xintercept = mean(flightsUA$dep_delay), color = "red")+
labs(title = "Histogram of Departure Delays", x = "Departure Delays(in minutes)", y = "Count")
```




```
mean(flightsUA$dep_delay)
```

```
## [1] 12.10607
```

```
ggplot(data = flightsUA, mapping = aes(x = dep_delay)) +
  geom_histogram(breaks=seq(0,105, by=5), color="white", fill="#440079") +
  facet_wrap(~ timeOfDay, nrow=4)
```




```
mean(flightsUA$dep_delay[flightsUA$timeOfDay == "evening"])
```

```
## [1] 22.76572
```

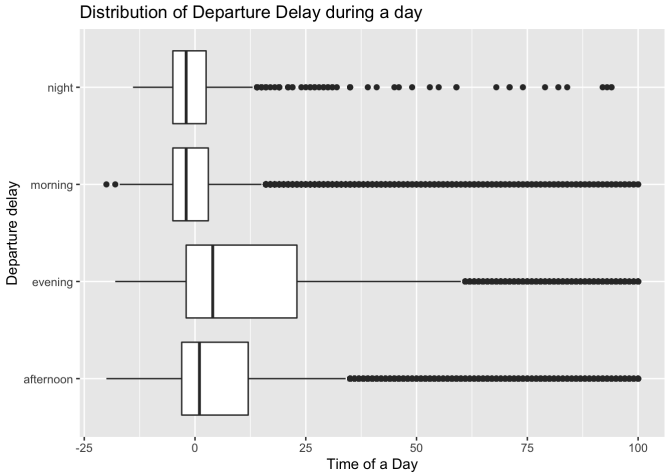
The data is right skewed and mean departure delay is around 10 minutes. We also can see there are some negative values which represent early departures as well as extremely long departure delays.

```
ggplot(data = flightsUA, aes(x = flightsUA$dep_delay, y = c(flightsUA$timeOfDay))) +
  geom_boxplot()+
  xlim(-20,100)+
  labs(x = "Time of a Day", y = "Departure delay", title = "Distribution of Departure Delay during a day")
```

```
## Warning: Use of `flightsUA$dep_delay` is discouraged. Use `dep_delay` instead.
```

```
## Warning: Use of `flightsUA$timeOfDay` is discouraged. Use `timeOfDay` instead.
```

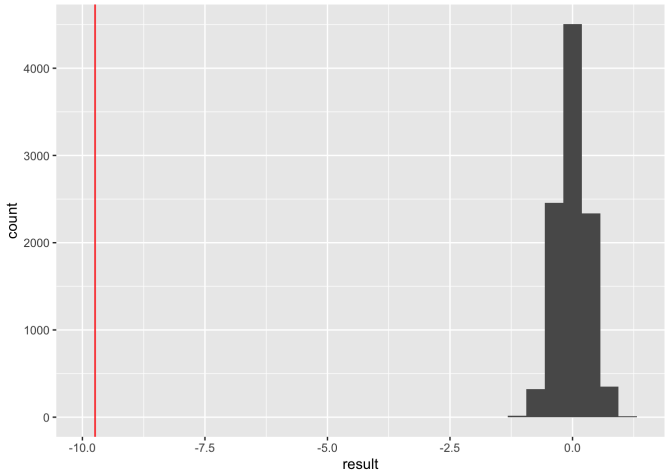
```
## Warning: Removed 1876 rows containing non-finite values (stat_boxplot).
```



From boxplots above we can see that median departure delay is roughly the same for all parts of a day, which is around 0 with some outliers. The evening has bigger IQR. Departures during morning and night have more early departures. To see if there is a pattern between departure delays versus time of a day we are going to perform two sided permutation tests. Our null hypothesis would be that there is no difference in means of departure delays. The alternative hypothesis is that there is real difference between times of a day.

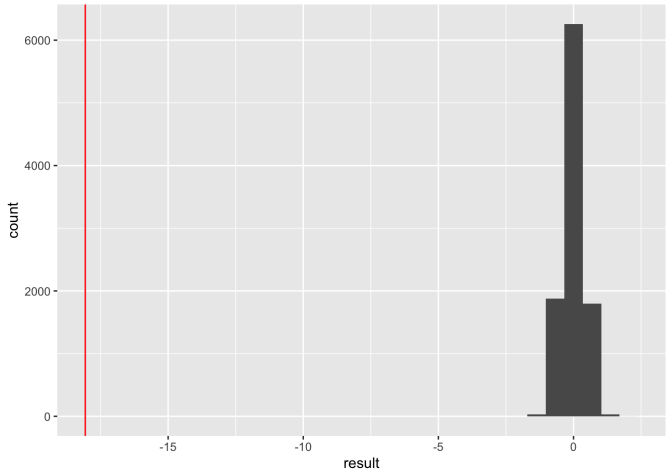
```
#N = number of simulations we will use
N <- 10^4-1

#create a blank vector to store the simulation results
result <- numeric(N)

#vector of the types of a day
vectorDay = c("morning", "afternoon", "evening", "night")

#loop through the types of a day and choose every time two of those
#calculate and store the observed difference in the sample
for(i in 1:length(vectorDay))
{
  for(j in 1:length(vectorDay)){
    if(j < 4 & i <= j){
      column1 = (vectorDay[i])
      column2 = (vectorDay[j+1])

      #reduce the data set to selected two seasons of a year
      reduced_flights <- flightsUA %>%
        filter(timeOfDay==column1 | timeOfDay==column2)

      #sample.size = the number of observations in our sample
      sample.size = nrow(reduced_flights)

      #group.1.size = the number of observations in the first group
      group.1.size = nrow(reduced_flights[reduced_flights$timeOfDay==column1,])

      #calculate the observed value
      observed <- mean(reduced_flights$dep_delay[reduced_flights$timeOfDay == column1])-
  mean(reduced_flights$dep_delay[reduced_flights$timeOfDay == column2])


      #use a for loop to cycle through values of k ranging from 1 to N
      for(k in 1:N)
        {
        index = sample(sample.size, size=group.1.size, replace = FALSE)
        result[k] = mean(reduced_flights$dep_delay[index])-
          mean(reduced_flights$dep_delay[-index])
        }

      #print the histograms
      print(ggplot(data=tibble(result), mapping = aes(x=result)) +
        geom_histogram() +
        geom_vline(xintercept = observed, color = "red"))

      #Calculate the p-value
      if(observed > 0){
        cat("The permutation for ", column1, " vs ", column2, ": ")
        print(p_value <- 2 * (sum(result >= observed) + 1) / (N + 1))
      }
      else{
        cat("The permutation for ", column1, " vs ", column2, ": ")
        print(p_value <- 2 * (sum(result <= observed) + 1) / (N + 1))
      }
    }
  }
}
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



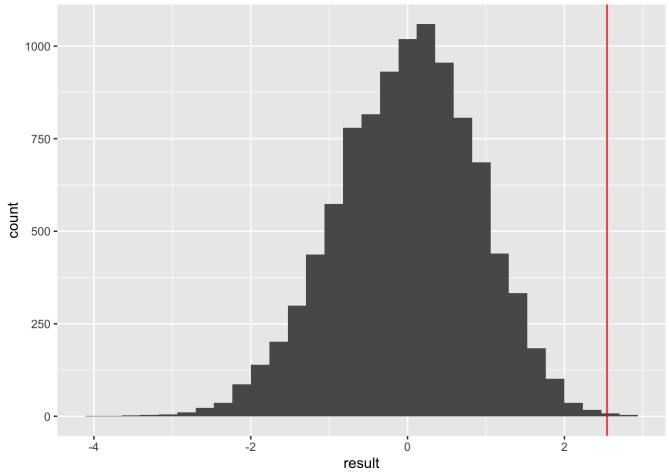

```
## The permutation for  morning  vs  afternoon : [1] 2e-04
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



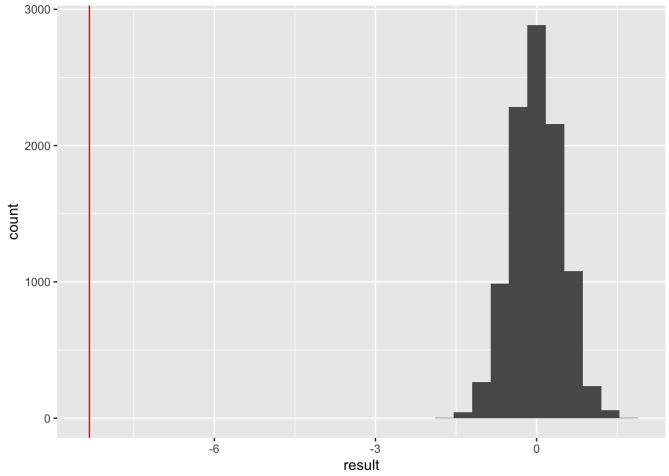

```
## The permutation for  morning  vs  evening : [1] 2e-04
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



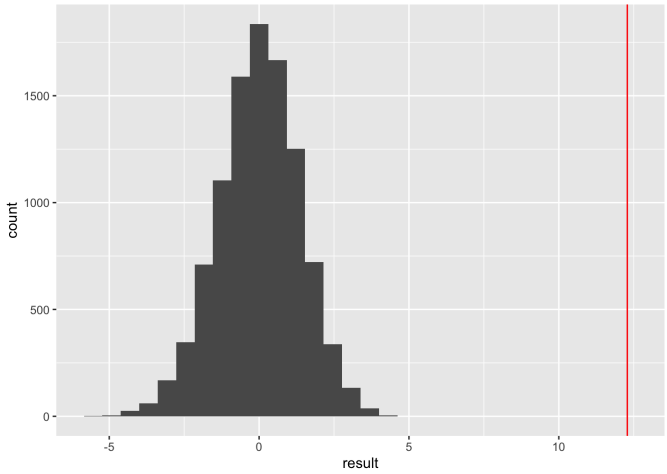

```
## The permutation for  morning  vs  night : [1] 0.0022
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



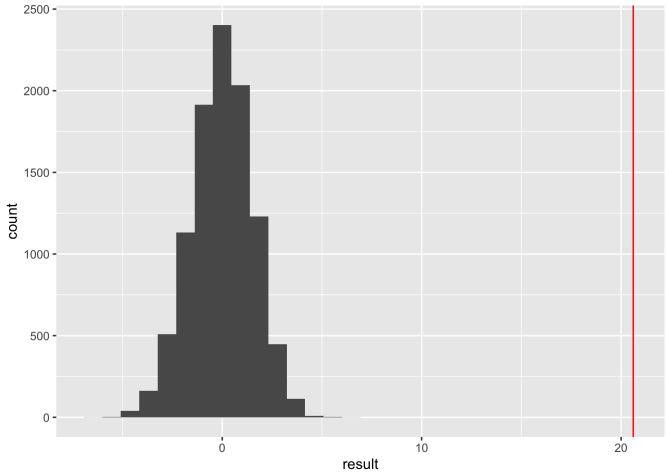

```
## The permutation for  afternoon  vs  evening : [1] 2e-04
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## The permutation for  afternoon  vs  night : [1] 2e-04
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



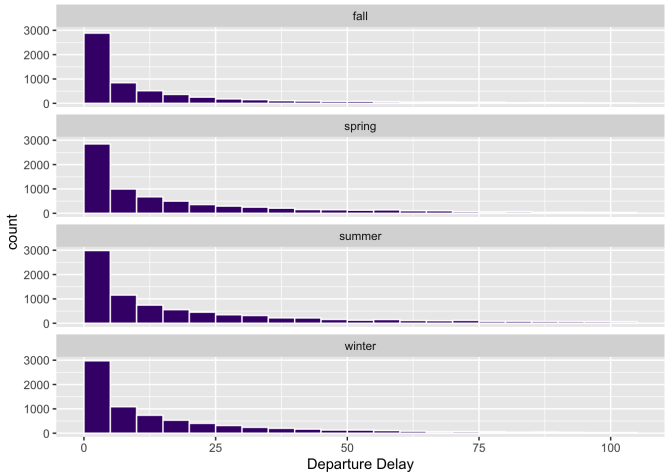

```
## The permutation for  evening  vs  night : [1] 2e-04
```

From the graphs we can see that simulated means are distributed normal around 0. The observed value is far from the simulated distribution which indicates that it is rare to get observed mean when we randomly categorize different types of a day. We observe that all our p values are small(0.0002) , however the one permutation case which involves morning vs night differs from others with slightly bigger p value = 0.0012. Since the p values are all less than 5%, we can conclude that there is significant difference in the means of departure delays between 4 times of a day thus we are rejecting null hypothesis in favor of the alternative hypothesis.

##2. Is there a pattern regarding departure delays versus time of year? Fidan

```
#create new variable which holds 4 seasons of a year
flightsUA<-mutate(flightsUA, timeOfYear = ifelse(month %in% 9:11, "fall",
                            ifelse(month %in% 3:5, "spring",
                            ifelse(month %in% 6:8, "summer","winter"))))
```

```
ggplot(data = flightsUA, mapping = aes(x = dep_delay)) +
  geom_histogram(breaks=seq(0,105, by=5), color="white", fill="#440079") +
  facet_wrap(~ timeOfYear, nrow=4)+
  xlab("Departure Delay")
```



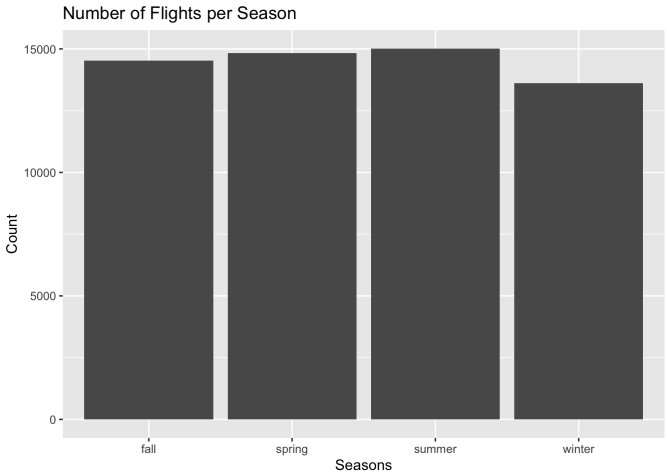
```
mean(flightsUA$dep_delay[flightsUA$timeOfYear == "summer"])
```

```
## [1] 17.54582
```

```
flightsUA %>%
  group_by(timeOfYear) %>%
  summarise(count = n())
```

```
## # A tibble: 4 × 2
##   timeOfYear count
##   <chr>      <int>
## 1 fall       14531
## 2 spring     14828
## 3 summer     15016
## 4 winter     13604
```

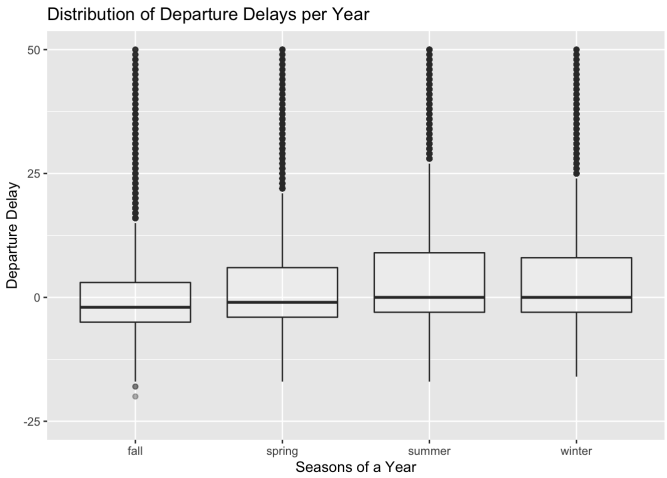
```
ggplot(data = flightsUA, mapping = aes(x = timeOfYear))+
  geom_bar()+
  labs(title = "Number of Flights per Season", x = "Seasons", y = "Count")
```



From bar charts we can see than the number of flights are almost the same for all 4 seasons.

```
ggplot(data = flightsUA, aes(x=timeOfYear, y=dep_delay))+
  geom_boxplot(alpha = 0.2) +
  ylim(-25,50) +
  labs(title = "Distribution of Departure Delays per Year", y = "Departure Delay", x = "Seasons of a Year")
```

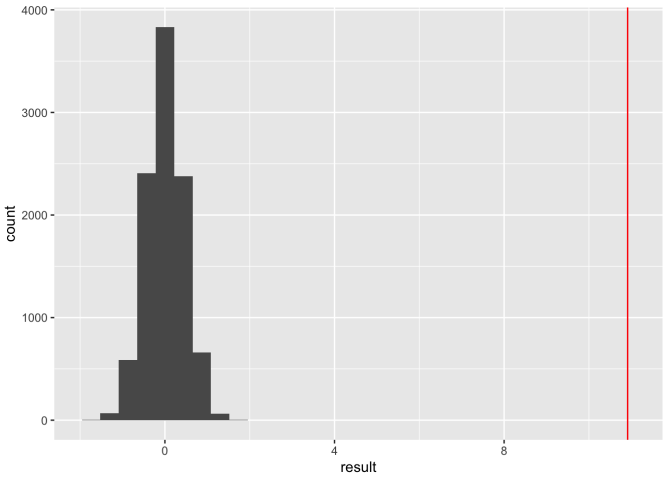
```
## Warning: Removed 4755 rows containing non-finite values (stat_boxplot).
```



The boxplot suggests that during fall and spring we have more early departures in contrast with summer and winter. The median is around 0 for summer and winter and slightly less for fall and spring. We can also see that outliers present on all of the seasons. To see if there is a pattern between departure delays versus time of a year we will devide a year into 4 seasons and perform two sided permutation tests. Our null hypothesis would be that there is no difference in means of departure delays. The alternative hypothesis is that there is a real difference between times of a day.

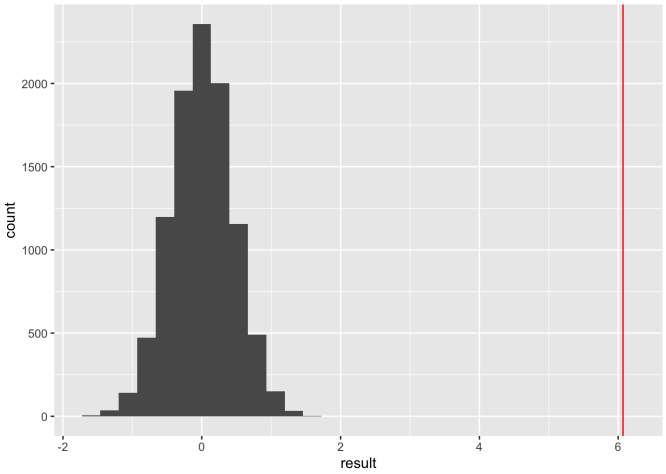
```
#N = number of simulations we will use
N <- 10^4-1

#create a blank vector to store the simulation results
result <- numeric(N)

#vector of the seasons
vectorYear = c("summer", "fall", "winter", "spring")

#loop through the seasons and choose each time two of those
#calculate and store the observed difference in the sample
for(i in 1:length(vectorYear))
{
  for(j in 1:length(vectorYear)){
    if(j < 4 & i <= j){
      column1 = (vectorYear[i])
      column2 = (vectorYear[j+1])

      #reduce the data set to those selected two seasons
      reduced_flights <- flightsUA %>%
        filter(timeOfYear==column1 | timeOfYear==column2)

      #sample.size = the number of observations in our sample
      sample.size = nrow(reduced_flights)

      #group.1.size = the number of observations in the first group
      group.1.size = nrow(reduced_flights[reduced_flights$timeOfYear==column1,])

      #calculate the observed value
      observed <- mean(reduced_flights$dep_delay[reduced_flights$timeOfYear == column1])-
  mean(reduced_flights$dep_delay[reduced_flights$timeOfYear == column2])


      #use a for loop to cycle through values of k ranging from 1 to N
      for(k in 1:N)
        {
        index = sample(sample.size, size=group.1.size, replace = FALSE)
        result[k] = mean(reduced_flights$dep_delay[index])-mean(reduced_flights$dep_delay[-index])

        }

      #print histograms
      print(ggplot(data=tibble(result), mapping = aes(x=result)) +
        geom_histogram() +
        geom_vline(xintercept = observed, color = "red"))

      #Calculate the p-value depending on observed value
      if(observed > 0){
        cat("The permutation for ", column1, " vs ", column2, ": ")
        print(p_value <- 2 * (sum(result >= observed) + 1) / (N + 1))
      }
      else{
        cat("The permutation for ", column1, " vs ", column2, ": ")
        print(p_value <- 2 * (sum(result <= observed) + 1) / (N + 1))
      }
    }
  }
}
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



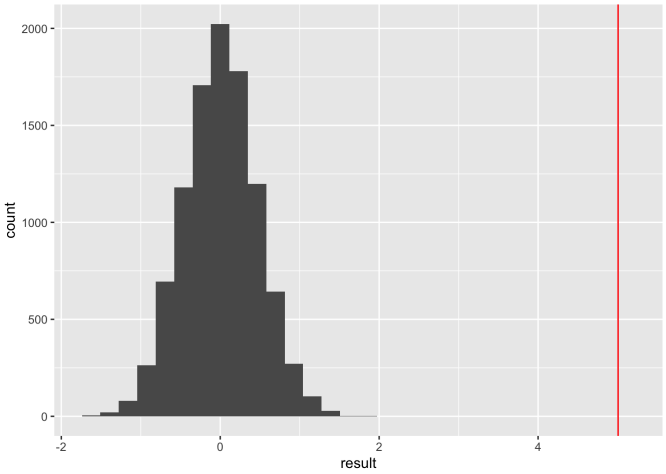

```
## The permutation for  summer  vs  fall : [1] 2e-04
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



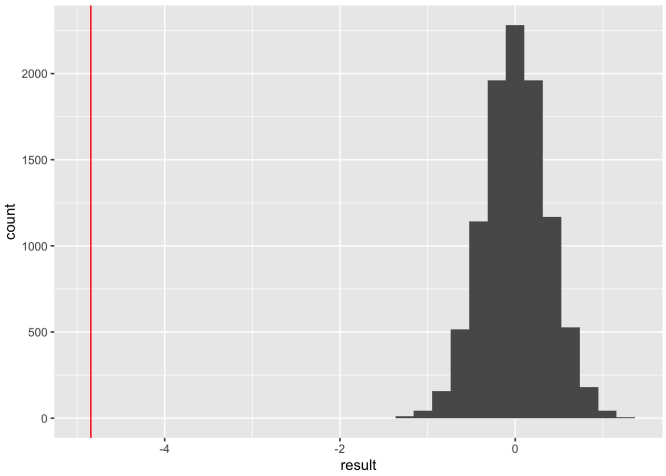

```
## The permutation for  summer  vs  winter : [1] 2e-04
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



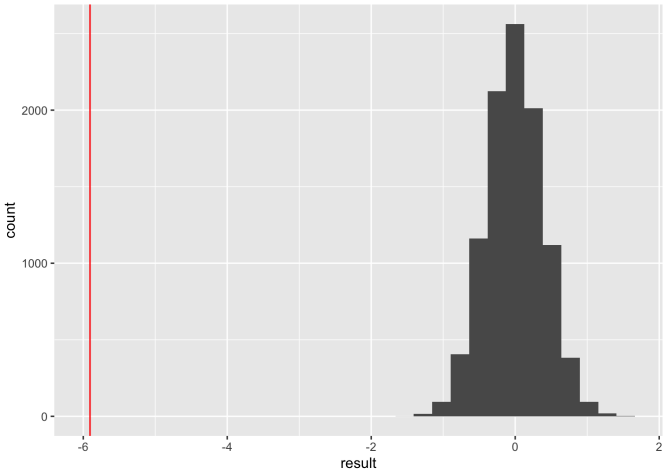

```
## The permutation for  summer  vs  spring : [1] 2e-04
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



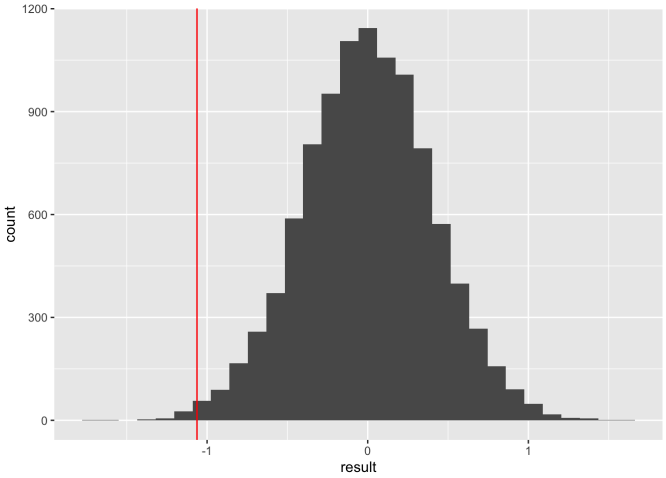

```
## The permutation for  fall  vs  winter : [1] 2e-04
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## The permutation for  fall  vs  spring : [1] 2e-04
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## The permutation for  winter  vs  spring : [1] 0.0088
```

From the graphs we can see that simulated means are distributed normal around 0. The observed value is far from the simulated distribution for all cases except winter vs spring, which indicates that it is rare to get observed mean when we randomly categorize different seasons of a year. We observe that all our p values are all small(0.0002) , however the one permutation case which involves winter vs spring differs from others with slightly bigger p value = 0.0094. Since the p values are all less than 5%, we can conclude that there is significant difference in the means of departure delays between 4 seasons of a year thus we are rejecting null hypothesis in favor of the alternative hypothesis.