Homework 4

Name: Arthur Viegas Eguia

I worked with: Deepak Bastola, Piper Dean (Stats Lab), Natalie Bax (Stats Lab)

Click the "Knit" button in RStudio to knit this file to a pdf.

Problem 1: flights

a.

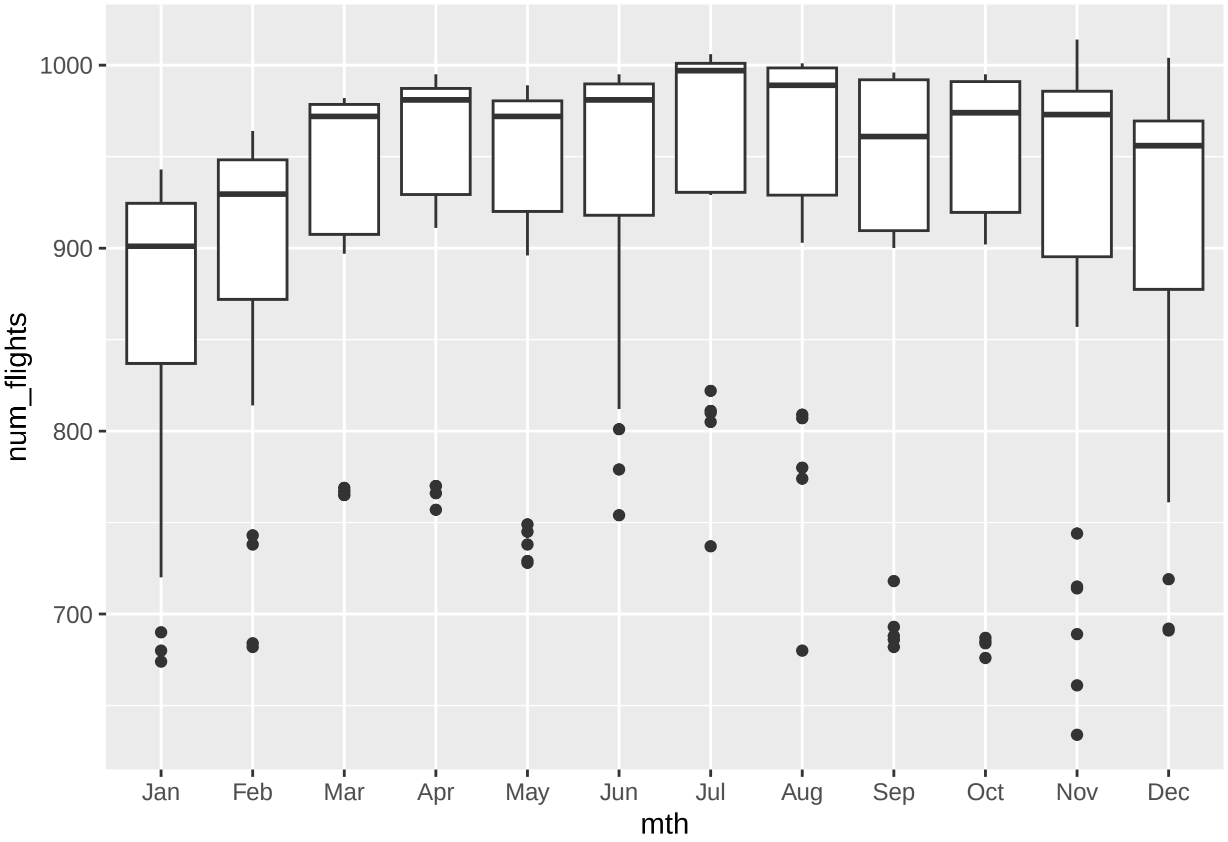
answer: The plane with tailnum N725MQ flew 575 times from New York

```
flight <- flights
mostTimesTailnum <- flights %>%
  group_by(tailnum) %>%
  summarise(totalCount = n()) %>%
  tidyr::drop_na() %>%
  arrange(desc(totalCount)) %>%
  slice(1)
mostTimesTailnum
## # A tibble: 1 x 2
##   tailnum totalCount
##   <chr>        <int>
## 1 N725MQ         575
```

b.

answer: In the summer, the months of July and August

```
flightsMonth <- flights %>%
  mutate(mth = month(time_hour, label = TRUE)) %>%
  group_by(day, mth) %>%
  summarise(num_flights = n())
ggplot(flightsMonth, aes(y = num_flights, x = mth)) + geom_boxplot()
```

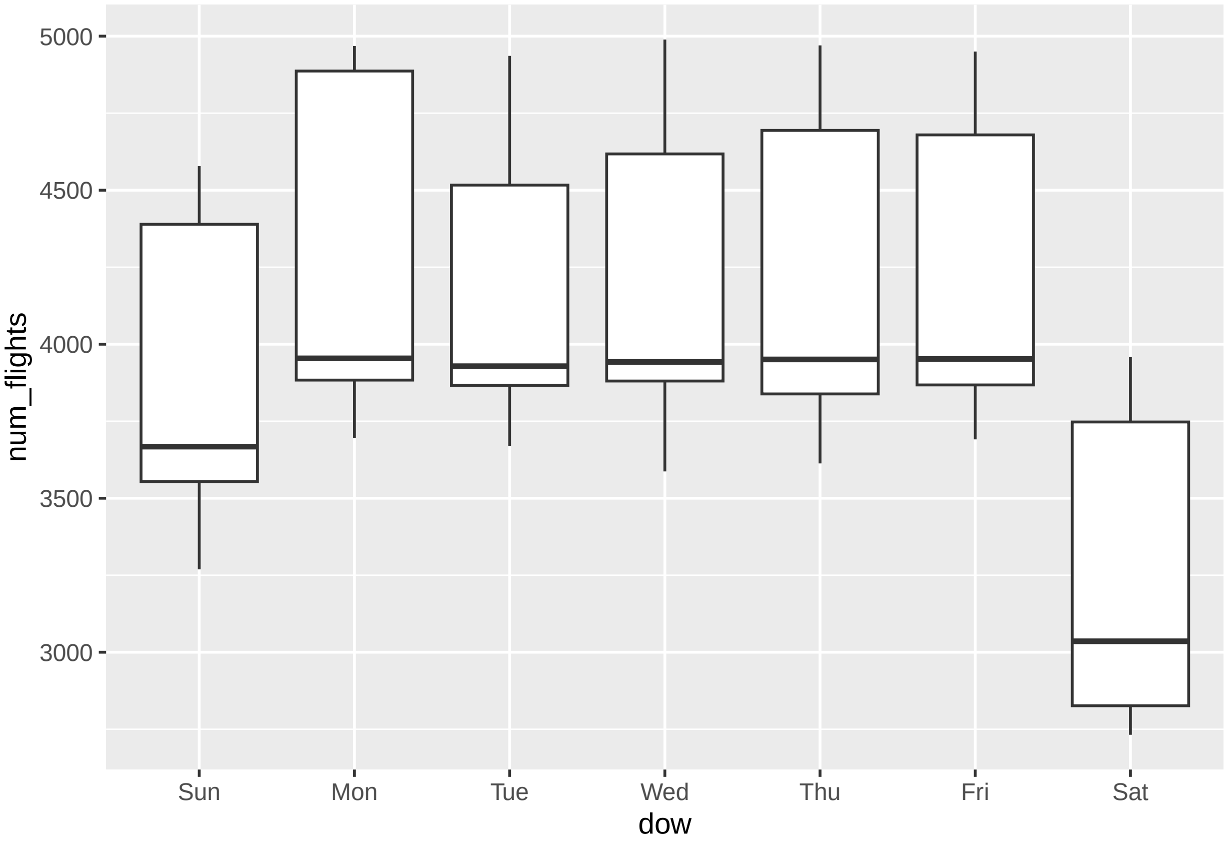


c.

answer: Saturday sees by far the fewest flights

```
flightsMonth <- flights %>%
  mutate(dow = wday(time_hour, label = TRUE)) %>%
  group_by(month, dow) %>%
  summarise(num_flights = n())

ggplot(flightsMonth, aes(y = num_flights, x = dow)) + geom_boxplot()
```



Problem 2: top destinations

```
top_dest <- flights %>%
  count(dest) %>%
  slice_max(n, n = 10)
```

a.

answer: The dimensions of this dataset are 141,145 × 20

```
top_dests_info <- top_dest %>%
  left_join(flight, by=c("dest"))
top_dests_info
## # A tibble: 141,145 x 20
##    dest      n  year month   day dep_time sched_dep_time dep_delay arr_time
##    <chr> <int> <int> <int> <int>    <int>          <int>     <dbl>    <int>
##  1 ORD   17283  2013     1     1      554            558        -4      740
##  2 ORD   17283  2013     1     1      558            600        -2      753
##  3 ORD   17283  2013     1     1      608            600         8      807
##  4 ORD   17283  2013     1     1      629            630        -1      824
##  5 ORD   17283  2013     1     1      656            700        -4      854
##  6 ORD   17283  2013     1     1      709            700         9      852
##  7 ORD   17283  2013     1     1      715            713         2      911
##  8 ORD   17283  2013     1     1      739            745        -6      918
##  9 ORD   17283  2013     1     1      749            710        39      939
## 10 ORD   17283  2013     1     1      828            830        -2      1027
## # i 141,135 more rows
## # i 11 more variables: sched_arr_time <int>, arr_delay <dbl>, carrier <chr>,
## #   flight <int>, tailnum <chr>, origin <chr>, air_time <dbl>, distance <dbl>,
## #   hour <dbl>, minute <dbl>, time_hour <dttm>
```

b.

answer: Median in code below

```
top_dests_info %>%
  group_by(dest) %>%
  mutate(dep_date = make_datetime(year = year,
                                   month = month,
                                   day = day,
                                   hour=hour,
                                   min = minute)) %>%
  arrange(dep_date) %>%
  mutate(dif_time = interval(lag(dep_date), dep_date) / dminutes(1)) %>%
  summarize(median_time=median(dif_time, na.rm = TRUE))
## # A tibble: 10 x 2
##    dest  median_time
##    <chr>       <dbl>
##  1 ATL            15
##  2 BOS            17
##  3 CLT            18
##  4 DCA            34
##  5 FLL            24
##  6 LAX            19
##  7 MCO            20
##  8 MIA            25
##  9 ORD            15
## 10 SFO            20
```

Problem 3: Energy

```
energy <- readr::read_csv("https://raw.githubusercontent.com/deepbas/statdatasets/main/energy.csv",
                    col_type = cols(
                     .default = col_double(),
                      Timestamp = col_datetime(format = ""),
                      dayWeek = col_factor(levels=c("Mon","Tues","Wed","Thurs","Fri","Sat","Sun"))
                     ))
dim(energy)
## [1] 35129    90
```

a.

answer: It contains 2880578 rows and 10 columns

```
names(energy)  # check variable names for use in pivot
##  [1] "Timestamp"                            "year"
##  [3] "month"                                "weekOfYear"
##  [5] "dayOfMonth"                           "dayWeek"
##  [7] "timeHour"                           "timeMinute"
##  [9] "100_Nevada_Street"                  "104_Maple_St."
## [11] "106_Winona_St."                     "Allen_House"
## [13] "Alumni_Guest_House/Johnson_House"   "Arboretum_Office"
## [15] "Art_Studios"                        "Benton_House"
## [17] "Berg_House"                         "Bird_House"
## [19] "Boliou_Memorial_Art_Bldg."          "Burton_Hall"
## [21] "Cassat_Hall_/_James_Hall"           "Center_for_Mathematics_&_Computing"
## [23] "Chaney_House"                       "Clader_House"
## [25] "College_Warehouse"                  "Cowling_Gym"
## [27] "Dacie_Moses_House"                  "Davis_Hall"
## [29] "Douglas_House"                      "Evans_Hall"
## [31] "Faculty_Club_/_Annex"               "Farm_House"
## [33] "Geffert_House"                      "Generator_Building"
## [35] "Goodhue_Hall"                       "Goodsell_Observatory"
## [37] "Gould_Memorial_Library"             "Grounds_Building"
## [39] "Headley_Cottage"                    "Headley_House"
## [41] "Henrickson_House"                   "Henry_House"
## [43] "Hill_House"                         "Hilton_House"
## [45] "Hoppin_House_(Alumni)"              "Hulings_Hall"
## [47] "Hunt_Cottage"                       "Huntington_House"
## [49] "James_Hall"                         "Jewett_House"
## [51] "Jones_House"                        "Laird_Hall"
## [53] "Laird_Stadium"                      "Language_&_Dining_Center"
## [55] "Leighton_Hall"                      "Main_Campus"
## [57] "Mudd_Hall_of_Science"               "Music_Hall"
## [59] "Musser_Hall"                        "Myers_Hall"
## [61] "Nourse_Hall"                        "Nutting_House"
## [63] "Olin_Hall_of_Science"               "Page_House_West"
## [65] "Parish_House_"                      "Parr_House"
## [67] "Pollock_House"                      "Prairie_Warehouse"
## [69] "Prentice_House"                     "Rayment_House"
## [71] "Recreation_Center"                  "Rice_House"
## [73] "Rogers_House"                       "Ryberg_House"
## [75] "Sayles-Hill"                        "Scoville_Hall"
## [77] "Seccombe_House"                     "Severance_Hall"
## [79] "Skinner_Memorial_Chapel"            "Sperry_House"
## [81] "Stimson_House"                      "Strong_House"
## [83] "Student_Townhouses"                 "Water_Tower"
## [85] "Watson_Hall"                        "Weitz_Center_for_Creativity"
## [87] "West_Gym"                           "Whittier_House"
## [89] "Willis_Memorial_Hall"               "Wilson_House"
energy_narrow <- energy %>%
  pivot_longer(
    names_to = "building",
    values_to = "energyKWH",
    cols = 9:90
  )
energy_narrow
## # A tibble: 2,880,578 x 10
##    Timestamp            year month weekOfYear dayOfMonth dayWeek timeHour
##    <dttm>              <dbl> <dbl>      <dbl>      <dbl> <fct>      <dbl>
##  1 2015-09-01 00:00:00  2015     9         35          1 Tues           0
##  2 2015-09-01 00:00:00  2015     9        35          1 Tues           0
##  3 2015-09-01 00:00:00  2015     9        35          1 Tues           0
##  4 2015-09-01 00:00:00  2015     9        35          1 Tues           0
##  5 2015-09-01 00:00:00  2015     9        35          1 Tues           0
##  6 2015-09-01 00:00:00  2015     9        35          1 Tues           0
##  7 2015-09-01 00:00:00  2015     9        35          1 Tues           0
##  8 2015-09-01 00:00:00  2015     9        35          1 Tues           0
##  9 2015-09-01 00:00:00  2015     9        35          1 Tues           0
## 10 2015-09-01 00:00:00  2015     9        35          1 Tues           0
## # i 2,880,568 more rows
## # i 3 more variables: timeMinute <dbl>, building <chr>, energyKWH <dbl>
```

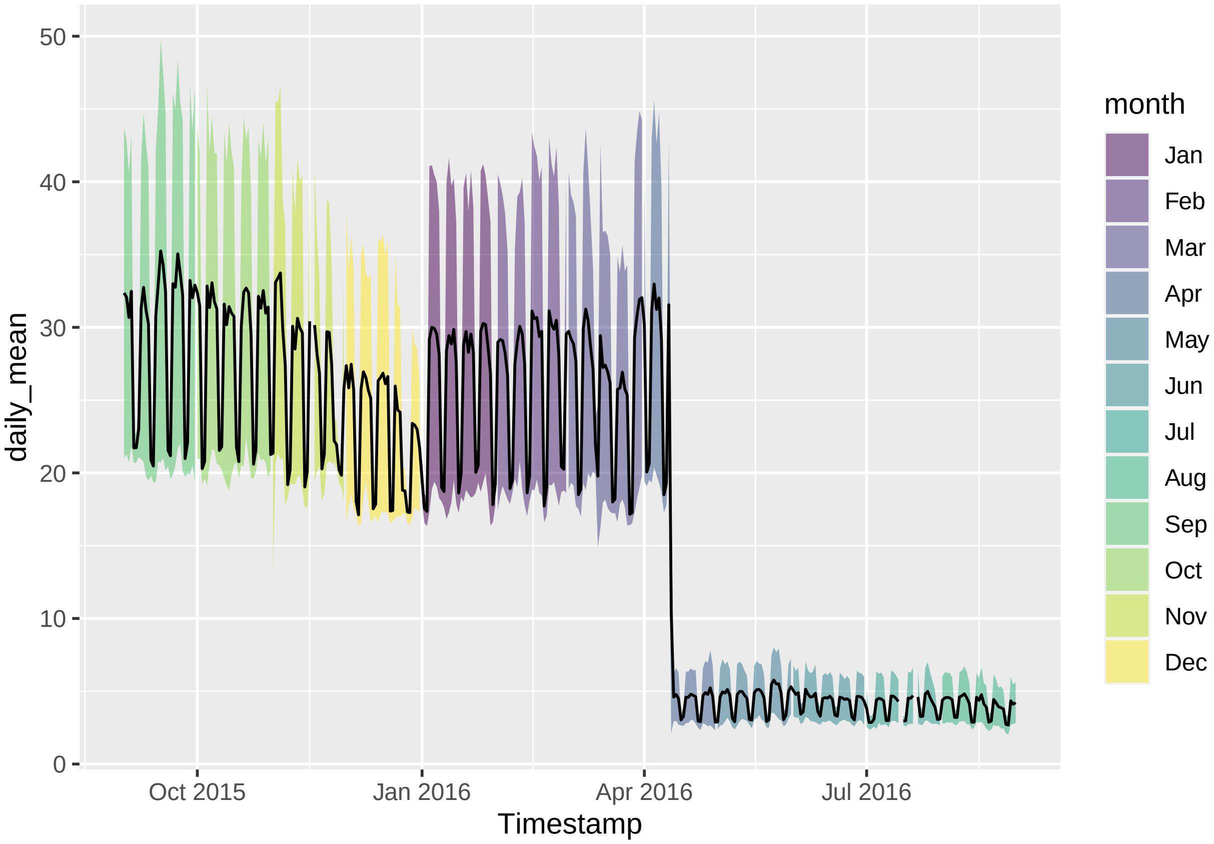
b.

```
lair_hall_data <- energy_narrow %>%
  mutate(Timestamp = date(Timestamp)) %>%
  group_by(Timestamp) %>%
  filter(building == "Laird_Hall") %>%
  summarise(daily_mean = mean(energyKWH),
            daily_standard_deviation = sd(energyKWH)) %>%
  mutate(month = month(Timestamp, label = TRUE))
lair_hall_data
## # A tibble: 366 x 4
##    Timestamp  daily_mean daily_standard_deviation month
##    <date>          <dbl>                    <dbl> <ord>
##  1 2015-09-01       32.4                    11.4  Sep
##  2 2015-09-02       32.1                    10.8  Sep
##  3 2015-09-03       30.7                     9.94 Sep
##  4 2015-09-04       32.5                    10.7  Sep
##  5 2015-09-05       21.7                     1.02 Sep
##  6 2015-09-06       21.7                     1.02 Sep
##  7 2015-09-07       23.0                     1.92 Sep
##  8 2015-09-08       31.3                    10.4  Sep
##  9 2015-09-09       32.8                    12    Sep
## 10 2015-09-10       31.2                    11.5  Sep
## # i 356 more rows
```

c.

answer: The consumption,both in daily mean and standard deviation seem to be very high until sometime in April. Then, by mid april both the mean and the standard deviation got much smaller.The mean went from values around 25 to 5, in mid April. Our interval (mean + and - the standard deviation) went from values in the 20s and 40s (before April), to values from around 3 to values around 8 after April.

```
ggplot(lair_hall_data, aes(x = Timestamp)) +
  geom_ribbon(aes(ymin = daily_mean - daily_standard_deviation,
                  ymax = daily_mean + daily_standard_deviation, fill = month),
              alpha = 0.5) + geom_line(aes(y = daily_mean))
```



d.

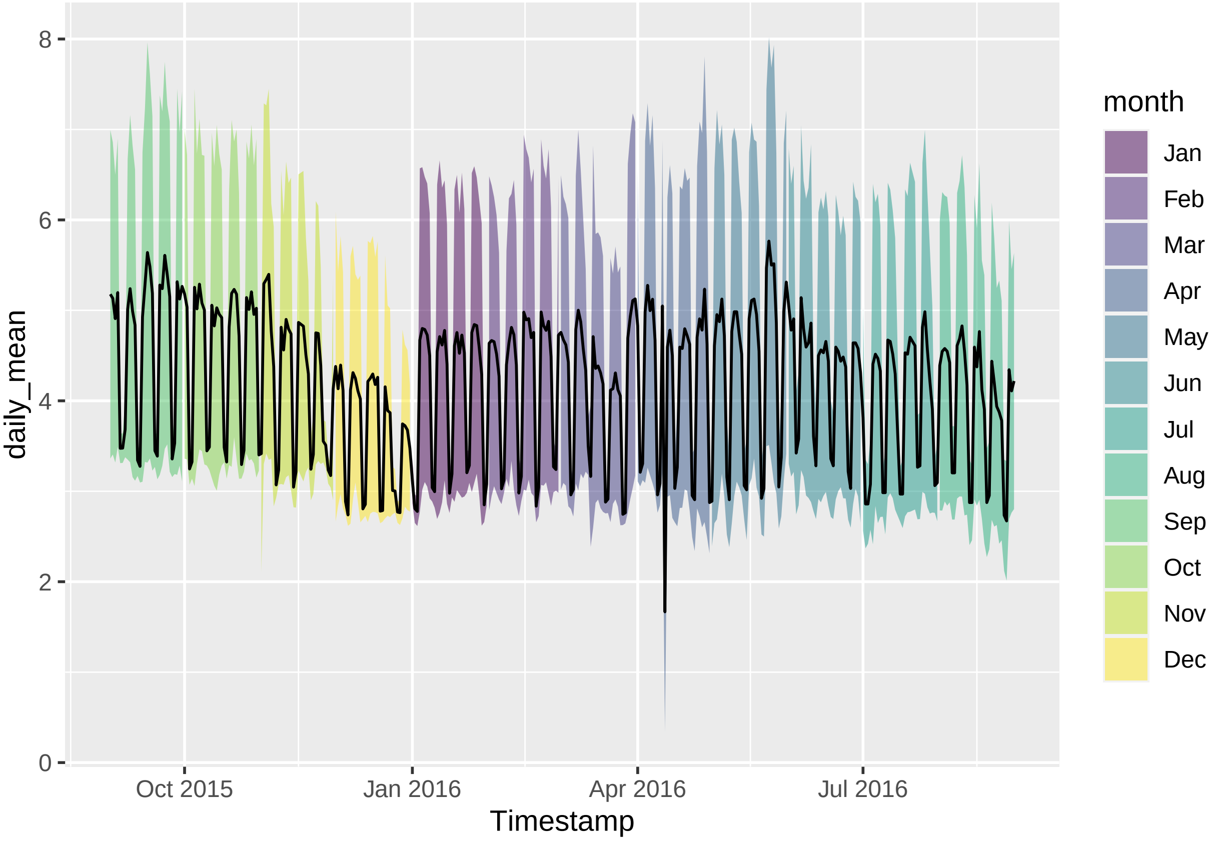
answer: It was April 12th.

```
day_adjusted <- lair_hall_data %>%
  drop_na()  %>%
  arrange(Timestamp) %>%
  filter(month == "Apr") %>%
  mutate(energy_diff = daily_mean - lag(daily_mean)) %>%
  arrange(energy_diff) %>%
  slice(1)
day_adjusted
## # A tibble: 1 x 5
##   Timestamp  daily_mean daily_standard_deviation month energy_diff
##   <date>          <dbl>                    <dbl> <ord>       <dbl>
## 1 2016-04-12       10.4                     8.30 Apr         -21.1
```

e.

answer: The data now follows a consistent pattern, considering a drop for Winter Break. Now, it seems that the pre-April readings are close to the post April readings.

```
laird_hall_data_adjusted <- lair_hall_data %>%
  drop_na() %>%
  arrange(Timestamp) %>%
  mutate(daily_mean = ifelse(Timestamp <= "2016-04-12", daily_mean * 0.16, daily_mean),
         daily_standard_deviation = ifelse(Timestamp <= "2016-04-12", daily_standard_deviation * 0.16, 
ggplot(laird_hall_data_adjusted, aes(x = Timestamp)) +
  geom_ribbon(aes(ymin = daily_mean + daily_standard_deviation,
                  ymax = daily_mean - daily_standard_deviation,
                  fill = month), alpha = 0.5) +
  geom_line(aes(y = daily_mean))
```



Problem 4: UN votes

```
unvotes <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/20
roll_calls <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data
issues <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/202
```

```
# Merge data frames
merged_data <- unvotes %>%
  left_join(roll_calls, by = "rcid", multiple = "all") %>%
  left_join(issues, by = "rcid", multiple = "all") %>%
  tidyr::drop_na(country, country_code, vote, issue, date) %>%
  mutate(vote = factor(vote))
```

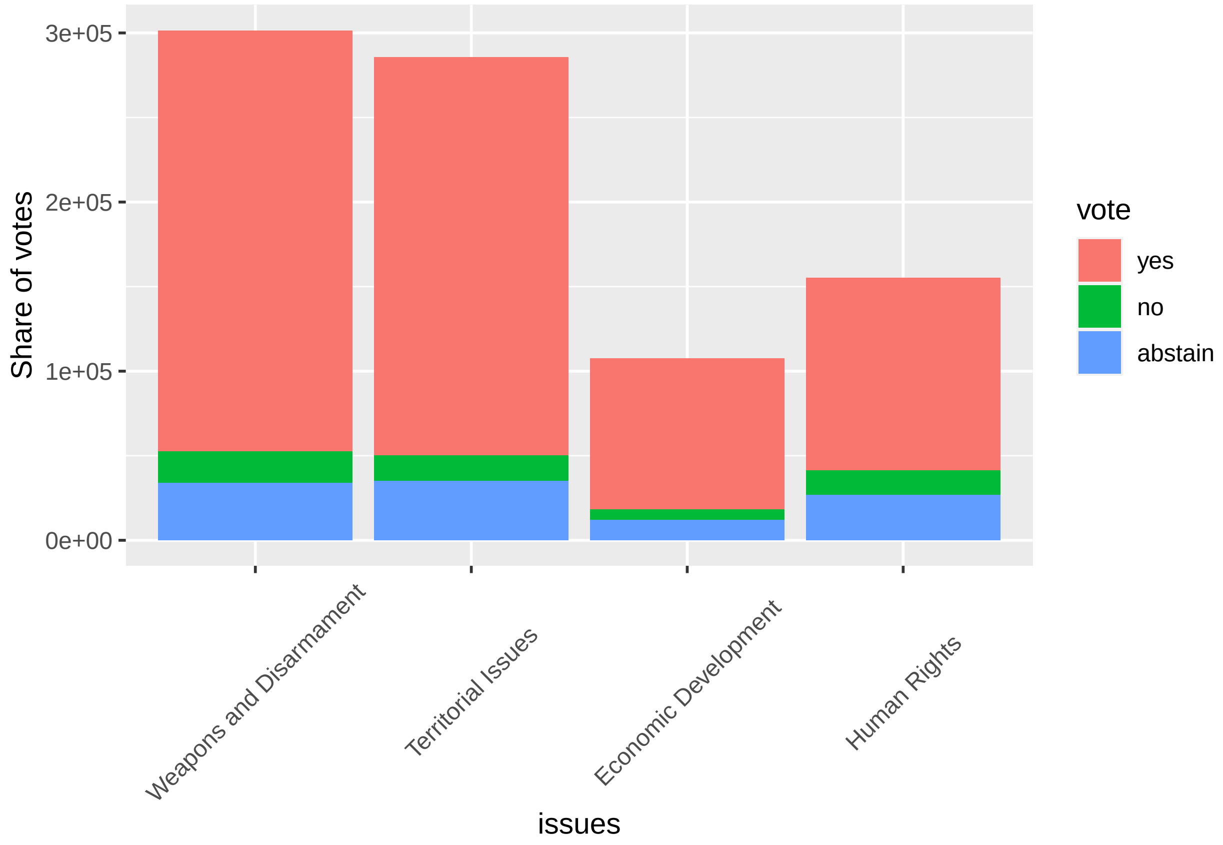
a.

```
unvotes_refactored <- merged_data %>%
  mutate(vote = fct_relevel(vote, c("yes", "no", "abstain"))) %>%
  arrange(vote)
unvotes_refactored
```

```
## # A tibble: 850,341 x 14
##     rcid country country_code vote  session importantvote date       unres amend
##    <dbl> <chr>   <chr>        <fct>   <dbl>         <dbl> <date>     <chr> <dbl>
##  1     6 Cuba    CU           yes         1             0 1946-01-04 R/1/~     0
##  2     6 Mexico  MX           yes         1             0 1946-01-04 R/1/~     0
##  3     6 Hondur~ HN           yes         1             0 1946-01-04 R/1/~     0
##  4     6 Nicara~ NI           yes         1             0 1946-01-04 R/1/~     0
##  5     6 Ecuador EC           yes         1             0 1946-01-04 R/1/~     0
##  6     6 Peru    PE           yes         1             0 1946-01-04 R/1/~     0
##  7     6 Chile   CL           yes         1             0 1946-01-04 R/1/~     0
##  8     6 Uruguay UY           yes         1             0 1946-01-04 R/1/~     0
##  9     6 Iran    IR           yes         1             0 1946-01-04 R/1/~     0
## 10     6 Lebanon LB           yes         1             0 1946-01-04 R/1/~     0
## # i 850,331 more rows
## # i 5 more variables: para <dbl>, short <chr>, descr <chr>, short_name <chr>,
## #   issue <chr>
```

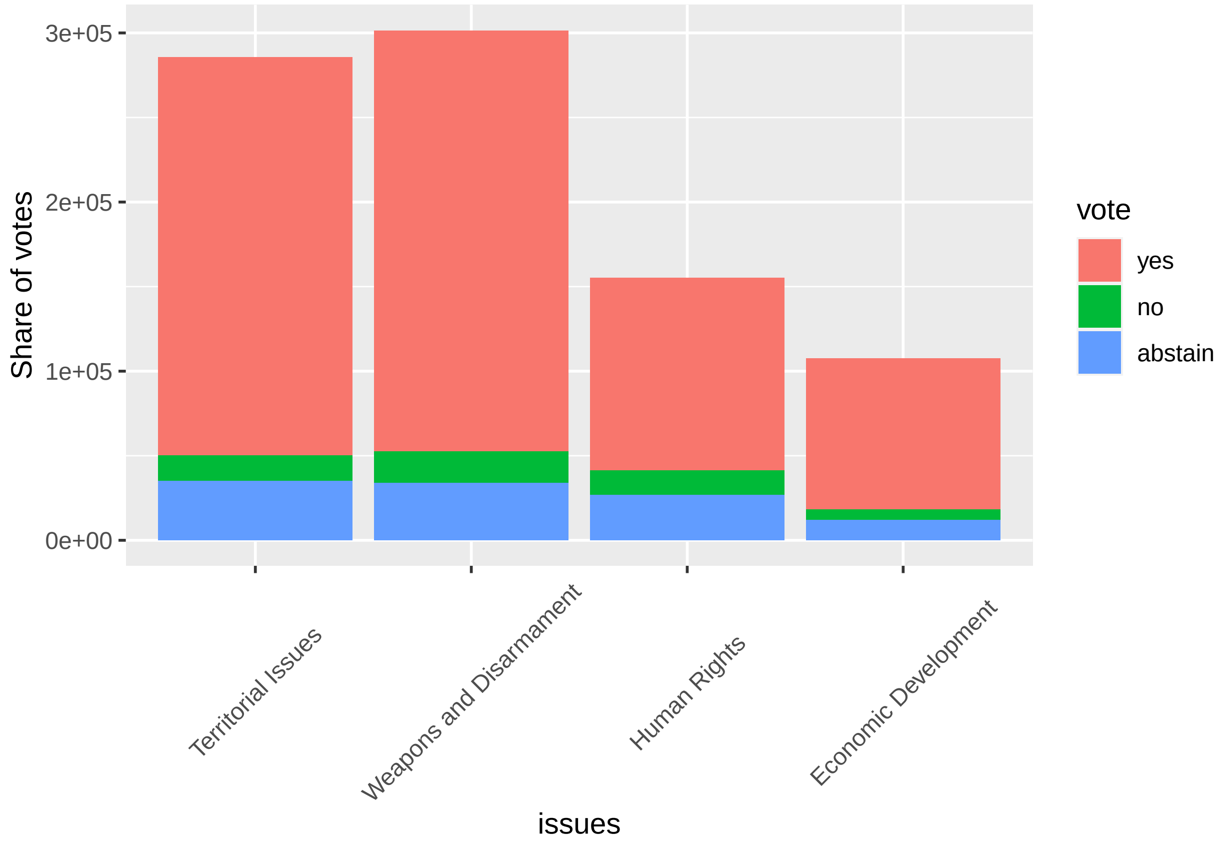
b.

```
merged_data %>% distinct(issue)
## # A tibble: 6 x 1
##   issue
##   <chr>
## 1 Human rights
## 2 Economic development
## 3 Colonialism
## 4 Palestinian conflict
## 5 Arms control and disarmament
## 6 Nuclear weapons and nuclear material
unvotes_refactored <- unvotes_refactored  %>%
  mutate(issue_factor = factor(issue),
         issue_category = fct_recode(issue_factor,
    "Territorial Issues" = 'Palestinian conflict',
    "Weapons and Disarmament" = 'Nuclear weapons and nuclear material',
    "Weapons and Disarmament"= 'Arms control and disarmament',
   "Territorial Issues"= 'Colonialism',
   "Economic Development" = 'Economic development', #This capitalizes Development
   "Human Rights" = 'Human rights'
      ))
ggplot(unvotes_refactored, aes(x = issue_category)) +
  geom_bar(aes(fill=vote)) +
  labs(y = "Share of votes", x="issues") +
  theme(axis.text.x = element_text(angle = 45, vjust = 0.5))
```



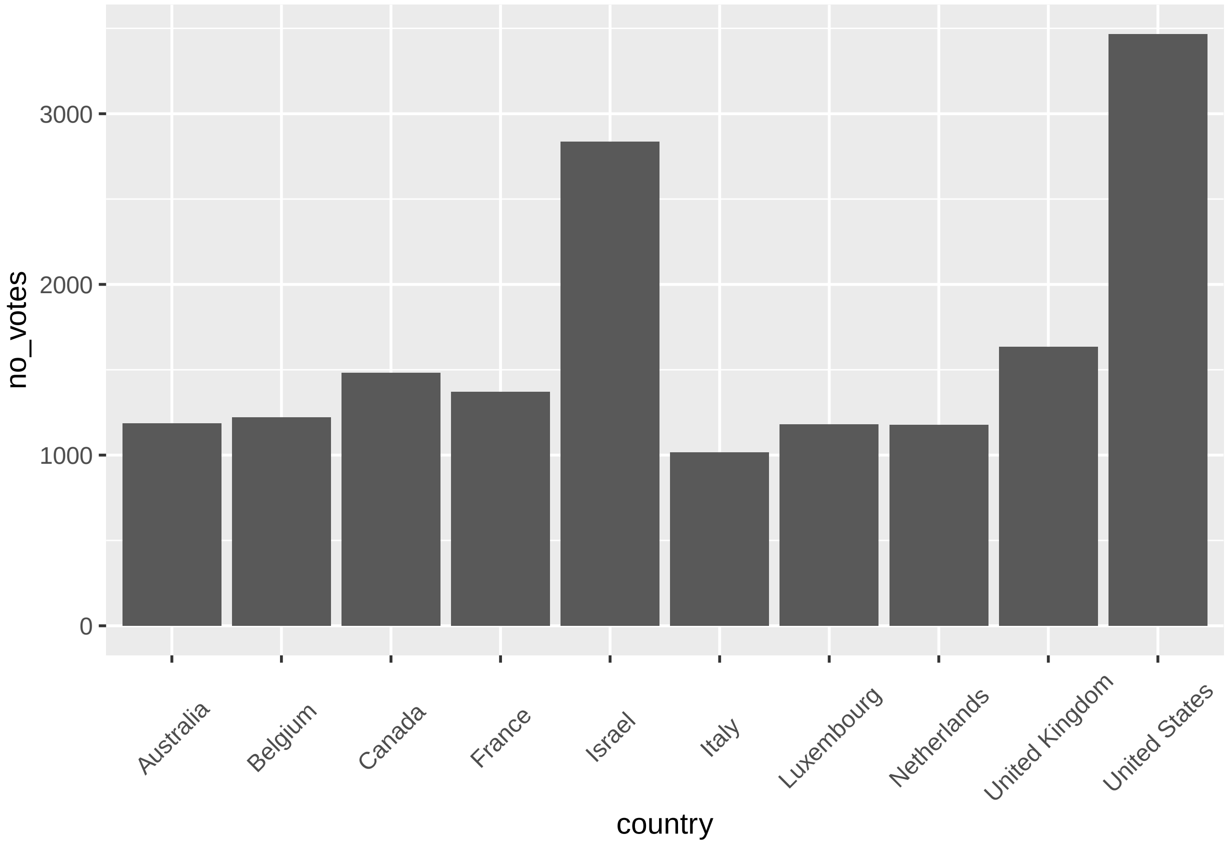
c.

```
unvotes_refactored <- unvotes_refactored %>%
  mutate(issue_category = fct_relevel(issue_category,
                                      c("Territorial Issues",
                                        "Weapons and Disarmament",
                                        "Human Rights",
                                        "Economic Development"))) %>%
  arrange(issue_category)
ggplot(unvotes_refactored, aes(x = issue_category)) +
  geom_bar(aes(fill=vote)) +
  labs(y = "Share of votes", x="issues") +
  theme(axis.text.x = element_text(angle = 45, vjust = 0.5))
```



d.

```
countries_no <- unvotes_refactored %>%
  group_by(country) %>% filter(vote == "no") %>%
  summarise(no_votes = n()) %>%
  arrange(desc(no_votes)) %>%
  slice(1:10)
countries_no
## # A tibble: 10 x 2
##    country        no_votes
##    <chr>             <int>
##  1 United States      3465
##  2 Israel             2836
##  3 United Kingdom     1633
##  4 Canada             1483
##  5 France             1369
##  6 Belgium            1220
##  7 Australia          1187
##  8 Luxembourg         1179
##  9 Netherlands        1176
## 10 Italy              1016
ggplot(countries_no,
       aes(x = country, y=no_votes)) +
  geom_col() +
  theme(axis.text.x = element_text(angle = 45, vjust = 0.5))
```



e.

```
unvotes_refactored <- unvotes_refactored %>%
  mutate(country_fact = factor(country),
         region = fct_collapse(country_fact,
           Americas = c("United States", "Canada", "Brazil", "Argentina", "Mexico"),
           Europe = c("United Kingdom", "France", "Germany", "Italy", "Spain"),
           Asia = c("China", "Japan", "India", "South Korea", "Russia"),
           'Middle East' = c("Iran", "Israel", "Saudi Arabia", "Turkey", "United Arab Emirates")
           ))
unvotes_refactored
## # A tibble: 850,341 x 18
##     rcid country country_code vote  session importantvote date       unres amend
##    <dbl> <chr>   <chr>        <fct>   <dbl>         <dbl> <date>     <chr> <dbl>
##  1    11 United~ US           yes         1             0 1946-02-05 R/1/~     0
##  2    11 Canada  CA           yes         1             0 1946-02-05 R/1/~     0
##  3    11 Cuba    CU           yes         1             0 1946-02-05 R/1/~     0
##  4    11 Domini~ DO           yes         1             0 1946-02-05 R/1/~     0
##  5    11 Hondur~ HN           yes         1             0 1946-02-05 R/1/~     0
##  6    11 Venezu~ VE           yes         1             0 1946-02-05 R/1/~     0
##  7    11 Ecuador EC           yes         1             0 1946-02-05 R/1/~     0
##  8    11 Peru    PE           yes         1             0 1946-02-05 R/1/~     0
##  9    11 Brazil  BR           yes         1             0 1946-02-05 R/1/~     0
## 10    11 Bolivia BO           yes         1             0 1946-02-05 R/1/~     0
## # i 850,331 more rows
## # i 9 more variables: para <dbl>, short <chr>, descr <chr>, short_name <chr>,
## #   issue <chr>, issue_factor <fct>, issue_category <fct>, country_fact <fct>,
## #   region <fct>
```