Lotto Data Analysis project

Samir Fidai

Loading Packages into R

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.3     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

## here() starts at C:/Users/samir/Documents/GitHub/Lotto Project/Lotto-Project

##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

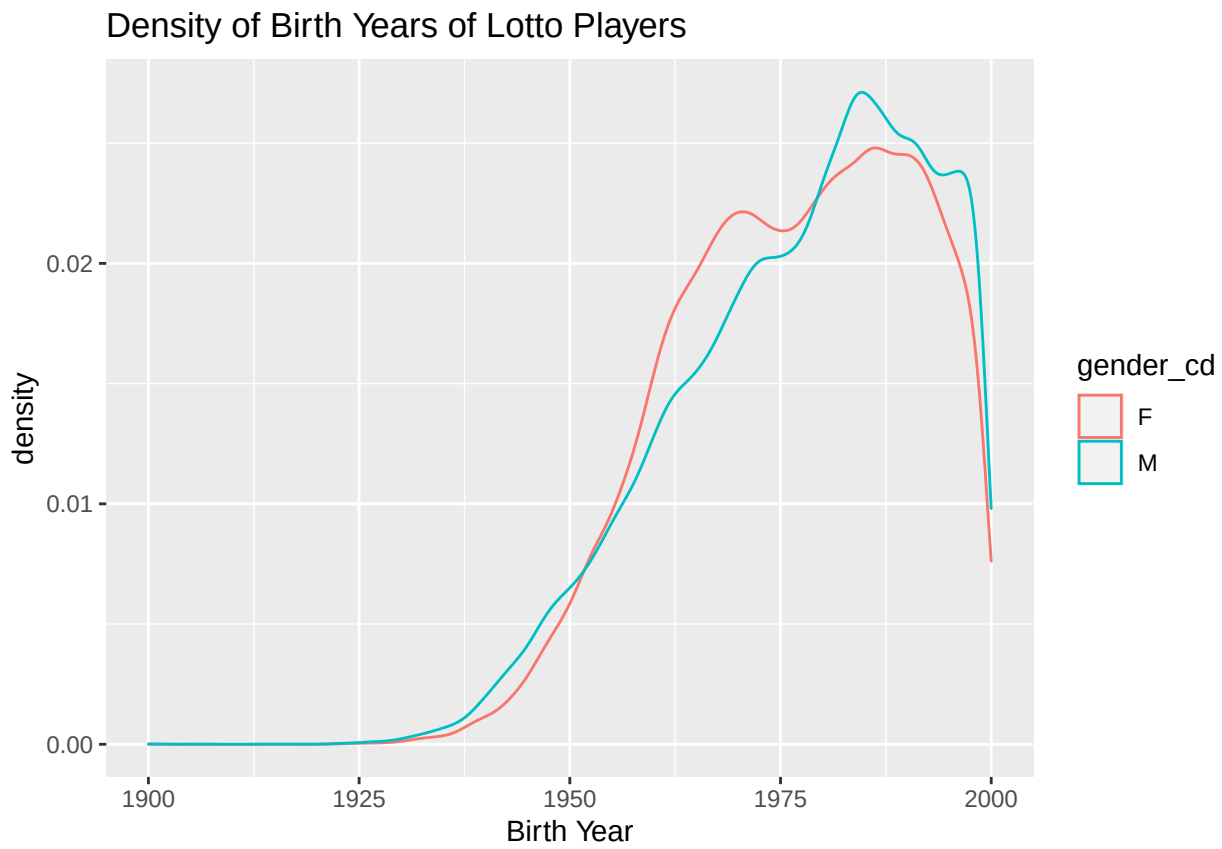
Loading the csv files into R. I switched the BOOKINGDATE variable in transactions from a character to a date format.

```
player <- read.csv("C://Users/samir/Documents/player.csv")
transaction_type <- read.csv("C://Users/samir/Documents/transaction_type.csv")
transaction <- read.csv("C://Users/samir/Documents/transaction.csv")
transaction <- transaction %>% mutate(BOOKINGDATE = lubridate::as_date(BOOKINGDATE))
```

For The first few chunks I'm just going to generate some charts to get a feel of the Player Demographics

The density plot below represents the distribution of the birth years. From the density plot we can get a feel of the distribution of age demographics

```
player %>%
  ggplot(aes(lubridate::year(birth_date), color = gender_cd)) +
  geom_density() +
  xlab("Birth Year") +
  ggtitle("Density of Birth Years of Lotto Players")
```

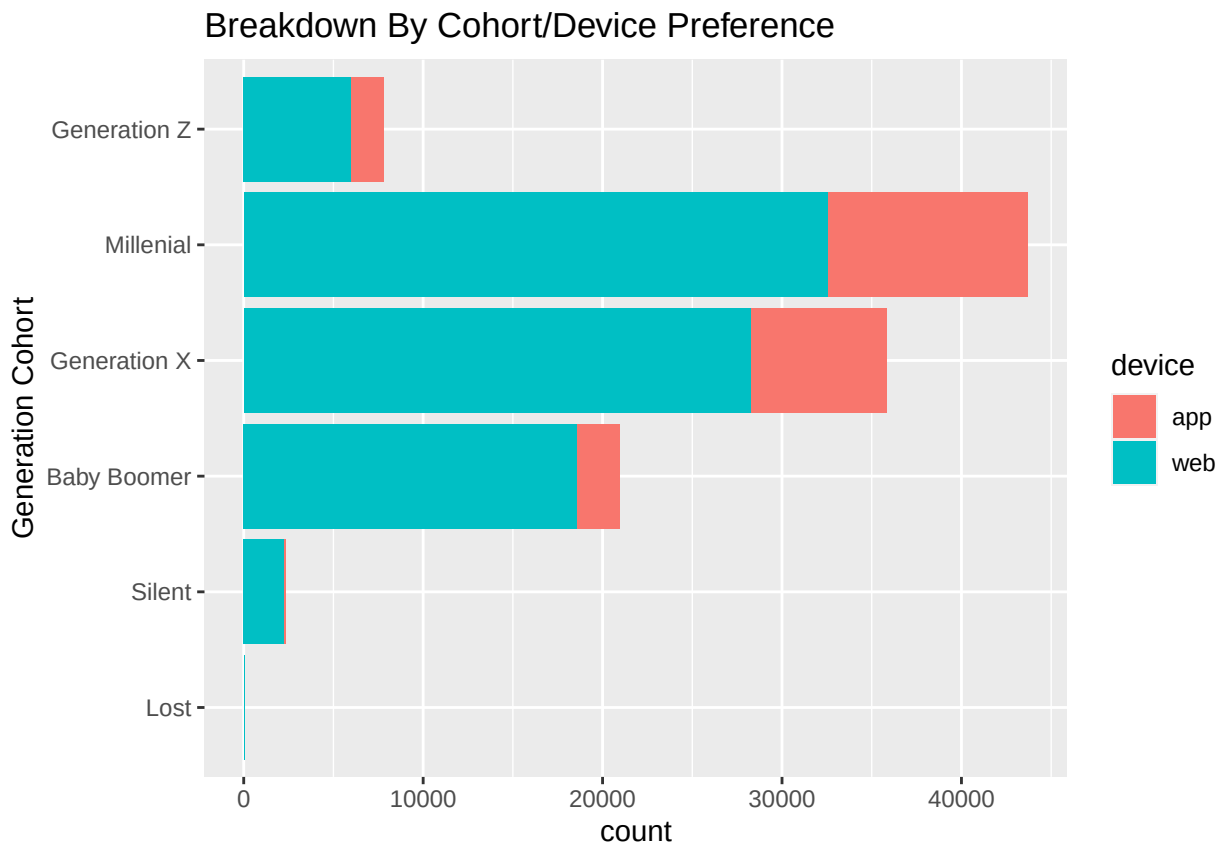


In this chart I aim to identify the preferred platforms each generational cohort uses.

```
levels <- c("Lost", "Silent", "Baby Boomer", "Generation X", "Millenial", "Generation Z")

player_by_cohort <- player %>%
  mutate(cohort = case_when(
    year(birth_date) < 1928 ~ "Lost",
    year(birth_date) %in% 1929:1945 ~ "Silent",
    year(birth_date) %in% 1946:1964 ~ "Baby Boomer",
    year(birth_date) %in% 1965:1980 ~ "Generation X",
    year(birth_date) %in% 1981:1996 ~ "Millenial",
    year(birth_date) >= 1997 ~ "Generation Z"
  )) %>%
  filter(!is.na(cohort))

player_by_cohort %>%
  ggplot(aes(factor(cohort, levels = levels), fill = device)) +
    geom_bar(stat = "count") +
    coord_flip() +
    xlab("Generation Cohort") +
    ggtitle("Breakdown By Cohort/Device Preference")
```



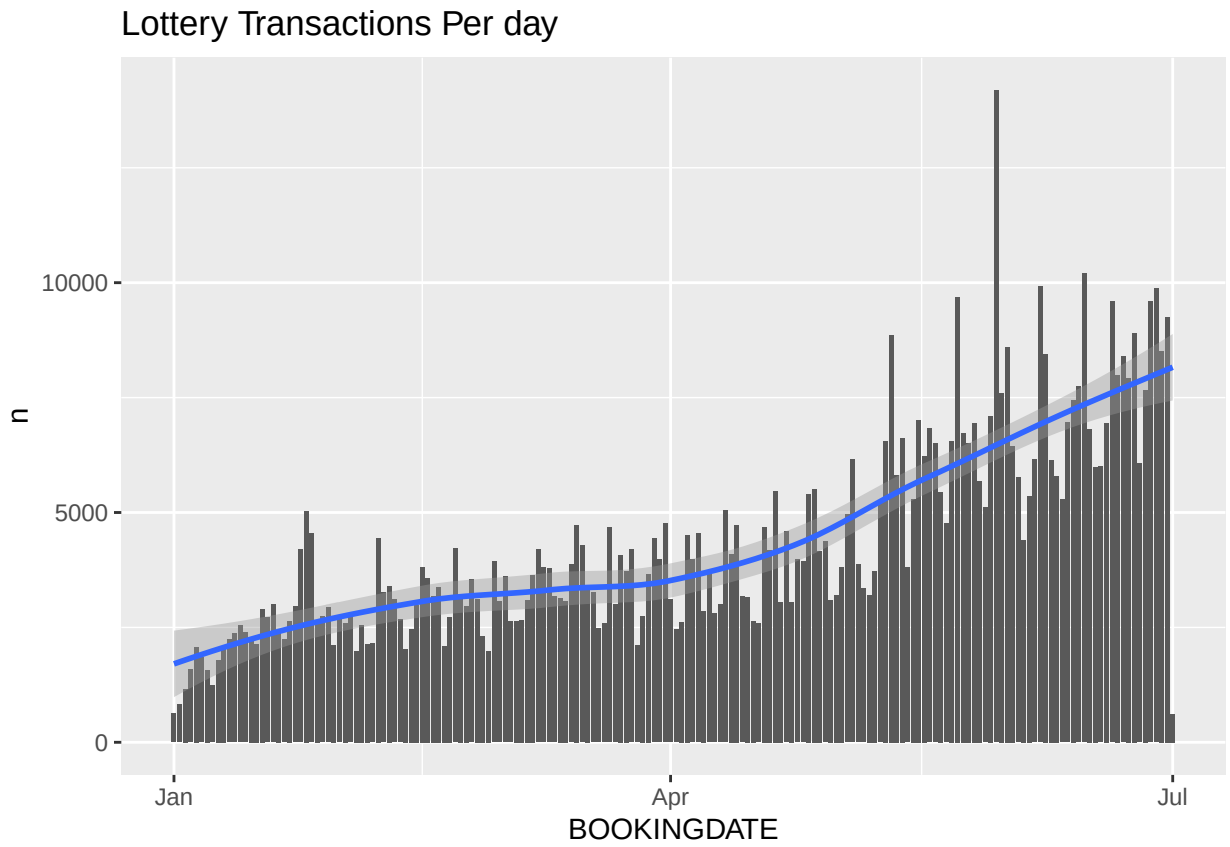
The two largest generational cohorts are Millennial and Generation X. Furthermore, although using the web platforms are dominant throughout the cohorts, the Millennial and Gen X cohorts do have frequent users of the applications.

The chart below aims to identify seasonal trends in when individuals want to purchase tickets

```
trans_per_date <- transaction %>%
  left_join(transaction_type, by = c("CLASSNAME" = "src_classname_txt")) %>%
  group_by(BOOKINGDATE) %>%
  filter(!is.na(TICKETTYPE), TICKETTYPE != "", trx_category_cd == "stake") %>%
  count(TICKETTYPE) %>%
  summarise(n = sum(n))

trans_per_date %>% ggplot(aes(x = BOOKINGDATE, y = n)) + geom_col() + geom_smooth() +
  ggtitle("Lottery Transactions Per day")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



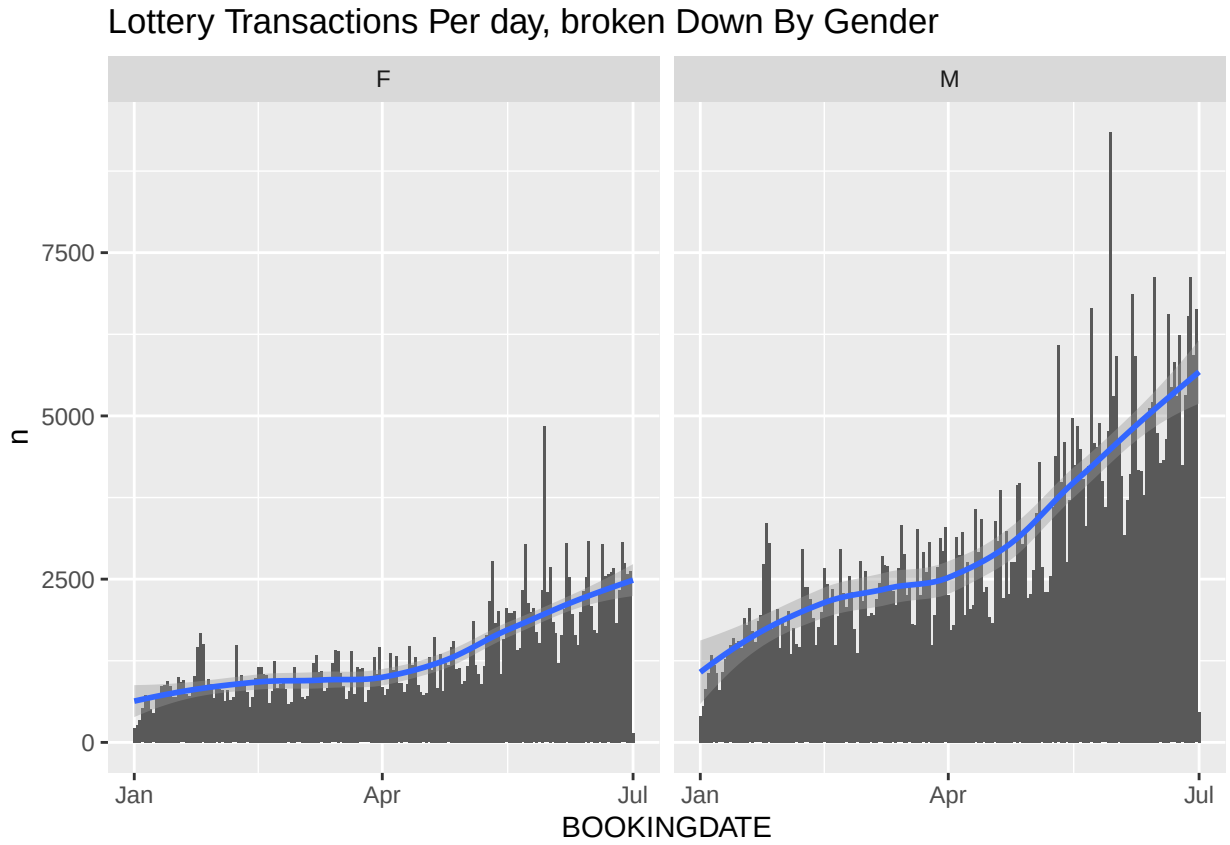
From the chart, we can see that as the season transitions into the summertime, an increased number of tickets get sold. Perhaps an ad campaign can be launched in late April or Early May with the goal of increasing lotto ticket purchases.

The following is the same chart from above but broken down by gender.

```
transaction %>%
  left_join(transaction_type, by = c("CLASSNAME" = "src_classname_txt")) %>%
  left_join(player, by = c("ACCOUNT_ID" = "src_account_id" )) %>%
  group_by(BOOKINGDATE, gender_cd) %>%
  filter(!is.na(TICKETTYPE), TICKETTYPE != "", trx_category_cd == "stake") %>%
  count(TICKETTYPE) %>%
  summarise(n = sum(n)) %>%
  ggplot(aes(x = BOOKINGDATE, y = n)) + geom_col() + geom_smooth() +
  ggtitle("Lottery Transactions Per day, broken Down By Gender") +
  facet_wrap(~gender_cd)
```

```
## `summarise()` has grouped output by 'BOOKINGDATE'. You can override using the `.groups` argument.
```

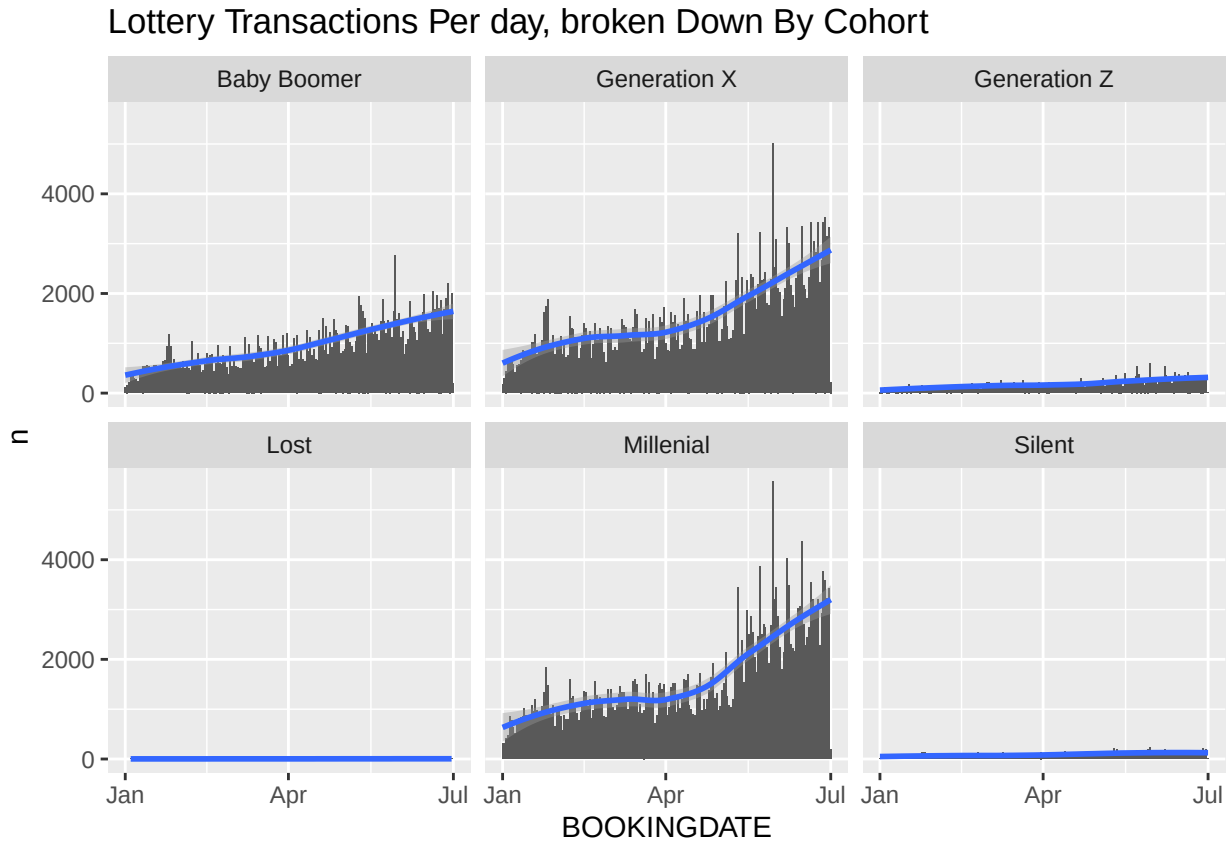
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
transaction %>%
  left_join(transaction_type, by = c("CLASSNAME" = "src_classname_txt")) %>%
  left_join(player_by_cohort, by = c("ACCOUNT_ID" = "src_account_id" )) %>%
  group_by(BOOKINGDATE, cohort) %>%
  filter(!is.na(TICKETTYPE), TICKETTYPE != "", trx_category_cd == "stake", !is.na(cohort)) %>%
  count(TICKETTYPE) %>%
  summarise(n = sum(n)) %>%
  ggplot(aes(x = BOOKINGDATE, y = n)) + geom_col() + geom_smooth() +
  ggtitle("Lottery Transactions Per day, broken Down By Cohort") +
  facet_wrap(~cohort)
```

```
## `summarise()` has grouped output by 'BOOKINGDATE'. You can override using the `.groups` argument.
```

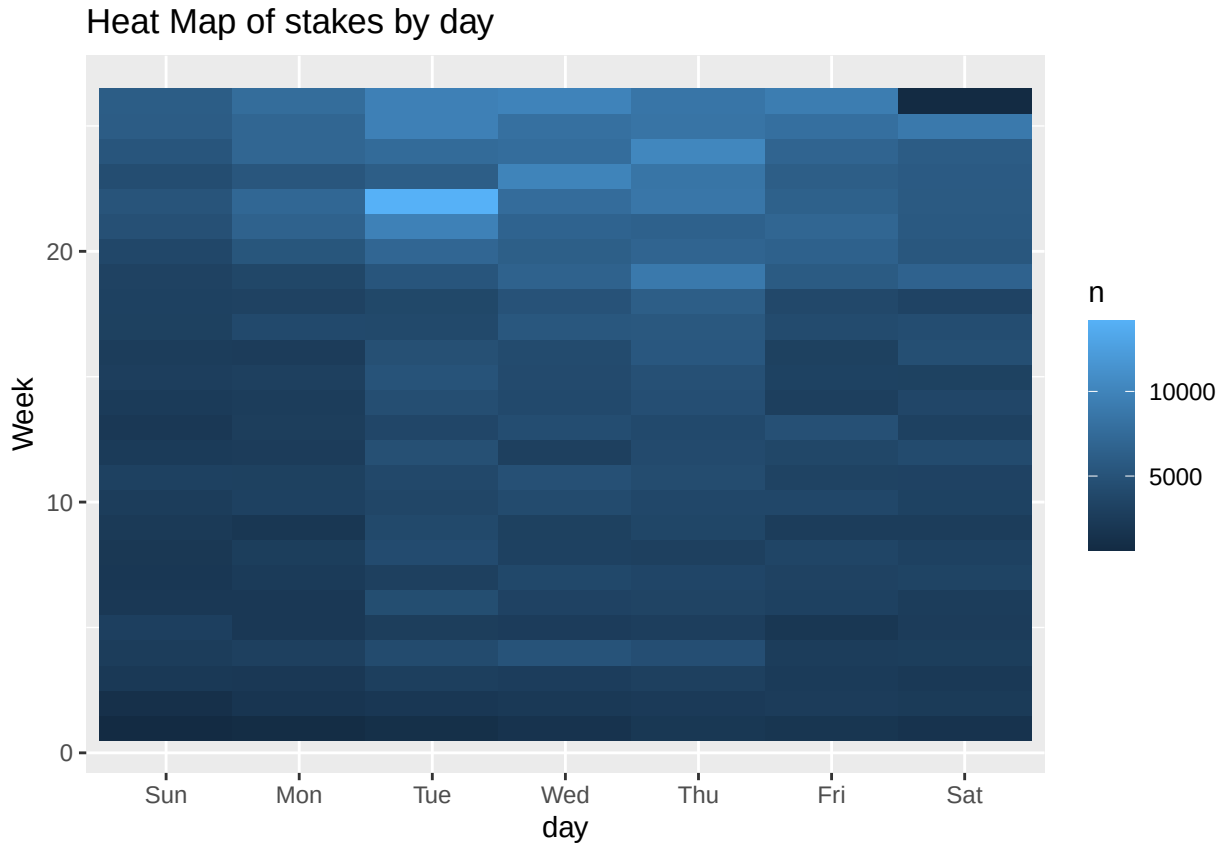
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



When breaking the stakes by cohort, there are some noticeable patterns. The baby boomer generation has more of a steady increase through the time period. The Generation Z cohort is more steady when it comes to creating stakes. Lastly, Generation X and the Millennial cohorts witness similar Springtime spikes in purchasing of stakes.

What I next wanted to do was identify patterns among the individual days. In order to visualize the patterns, I opted to use a heat map.Using lubridate I pulled the days of the week out from the dates, and I put the Week on the y-axis.

```
trans_per_date %>%
  mutate(day = wday(BOOKINGDATE, label = TRUE),
         month = month(BOOKINGDATE, label = TRUE)) %>%
  ggplot(aes(x = day, y = week(BOOKINGDATE), fill = n)) +
  geom_tile() +
  ylab("Week") +
  ggtitle("Heat Map of stakes by day")
```



From the heat map we can see that stake purchasing activity is at its lowest during the weekends. However, the purchasing of stakes picks up during the middle of the week. The lighter the color the more stakes were placed on that day.

The next question I wanted to ask was "Are purchasing stakes actually worth it?" To answer this question, I grouped the Account Numbers along with their Ticket types. I summed up the total stakes for each ticket type they bought. In another table I calculated the winnings by filtering out the "wins" and subtracting the POSTBALANCE from the Pre Balance. After joining the tables I calculated the difference by subtracting the winnings by the stakes. A positive stake-win difference would indicate that the Account owner actually made a profit from buying stakes for that specific ticket type. A negative value indicates an account owner lost money on the ticket. The relationship was visualized in a scatter plot. Also in order to avoid confusion, I strictly only used accounts who used the Euro as their currency.

```
stake <- transaction %>%
  left_join(transaction_type, by = c("CLASSNAME" = "src_classname_txt")) %>%
  left_join(player, by = c("ACCOUNT_ID" = "src_account_id")) %>%
  filter(trx_category_cd == "stake", currency_cd =="EUR  ") %>%
  mutate(stakes = PREBALANCE - POSTBALANCE) %>%
  group_by(ACCOUNT_ID, TICKETTYPE) %>%
  summarise(stakes = sum(stakes))
```

```
## `summarise()` has grouped output by 'ACCOUNT_ID'. You can override using the `.groups` argument.
```

```
win <- transaction %>%
  left_join(transaction_type, by = c("CLASSNAME" = "src_classname_txt")) %>%
  left_join(player, by = c("ACCOUNT_ID" = "src_account_id")) %>%
  mutate(winnings = POSTBALANCE - PREBALANCE) %>%
```

```
  filter(trx_category_cd == "win",currency_cd == "EUR  ") %>%
  group_by(ACCOUNT_ID,TICKETTYPE) %>%
  summarise(winnings = sum(winnings))
```

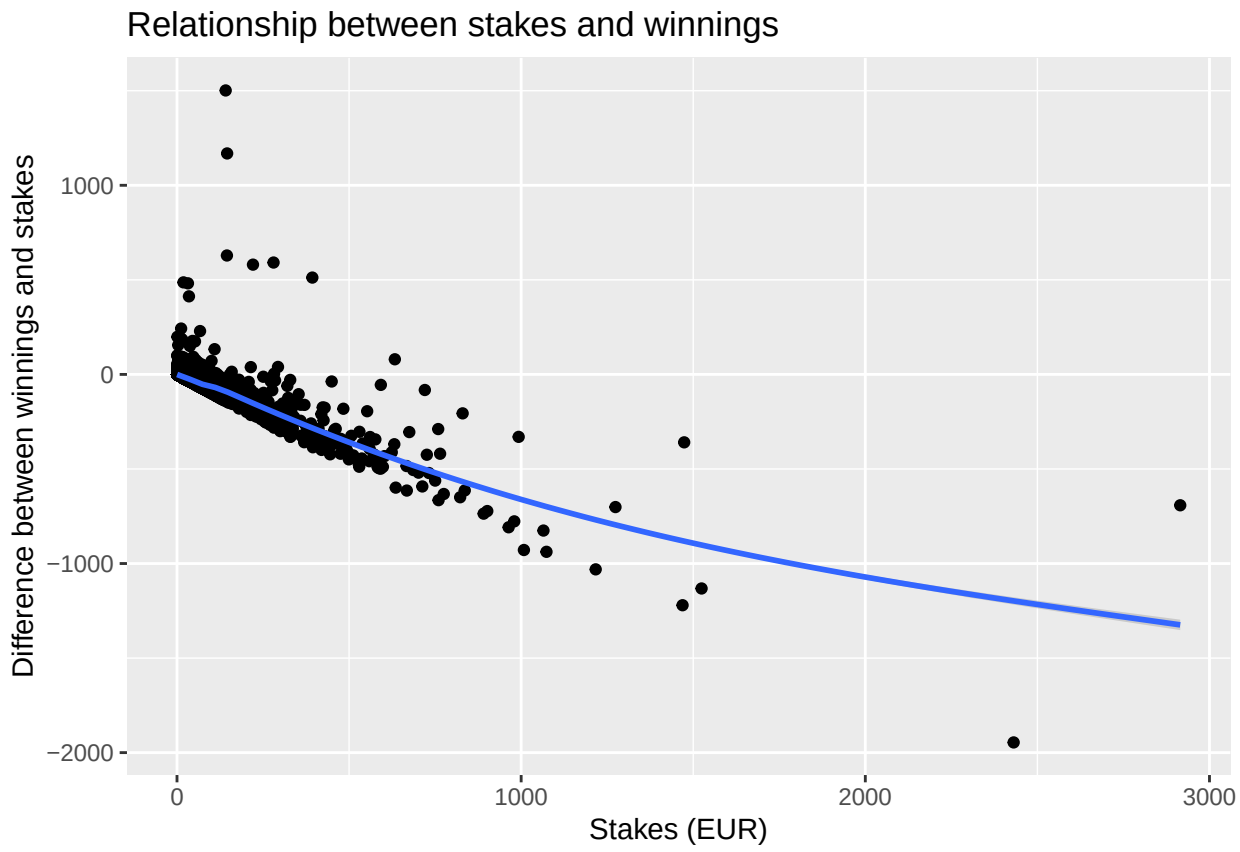
```
## `summarise()` has grouped output by 'ACCOUNT_ID'. You can override using the `.groups` argument.
```

```
combined <- stake %>%
  full_join(win) %>%
  mutate(winnings = replace_na(winnings, 0),
         `stake-win diff` = winnings - stakes)
```

```
## Joining, by = c("ACCOUNT_ID", "TICKETTYPE")
```

```
combined %>%
  filter(`stake-win diff` <= 10000) %>%
  ggplot(aes(x = stakes, y = `stake-win diff`)) +
  geom_point() +
  geom_smooth() +
  ylab("Difference between winnings and stakes") +
  xlab("Stakes (EUR)") +
  ggtitle("Relationship between stakes and winnings")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



The scatter plot along with its accompanying smooth indicates that the more an account purchases more of

a lotto ticket stake, their chances of them either winning or breaking even on the lotto ticket gets reduced less and less.