Final_Project

Neaha_Bijo_400433756

2023-04-10

Introduction

The data used in the project was obtained from Kaggle. Data will be taken from 1950-2022 (excluding the shortened 2017 season) for race results. This project focuses on the safety concerns in Formula 1. This covers the deployment of safety cars, red-flags, pole positions being safe to win the race. As safety cars deployments and red flags are always blamed on drivers and teams I want to try and find the other factors affecting the safety of the race.

- These are the links from where I sourced my datasets
  - Race Data
  - Race Events

I will ask the following questions:

1. Do drivers on pole position in qualifying always win the race in the Grand Prix? If not then why do they fight for it?
   - This is used to find out why drivers and teams always aim for first finish in qualifying even though the weekend winner is determined by the driver and team who wins in the Grand Prix.
2. How does the safety car deployments in the middle of the race change the situation of the race?
   - Safety cars a re deployed for various reasons but some teams and drivers are not happy when this occurs.
3. Which circuits have the most events going on? Is the track safe for a race to be on the calendar in the future?
   - Finding out which circuits have the most events (red flags, accidents etc) can help determine if the circuit is safe to race as cars are 2 times are big as they were in the 90s. Theoretically once a circuit is determined dangerous by FIA regulations that track should not be on the race calender and be either discontinued or renovated.

Data Wrangling Plan

- I have not included all the DWP of the files to save space.
- The other files I have tidyed are circuits.csv, lapTimes.csv, pitStops.csv, qualifying.csv, red_flags.csv.
- I have not used the maggrittr operator for the convenience of accessing datasets before tidying.

GP Result

Iteration 1

Phase 1

1. Read the csv file into R
2. Make column names lowercase
3. Determine if the data is Tidy and if not fix it
4. Identify uids
5. Drop unnecessary columns

Phase 2

```
##1.
results_tib <- read.csv("./Data/results.csv", ) %>% glimpse()
```

```
## Rows: 23,777
## Columns: 18
## $ resultId        <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,~
## $ raceId          <int> 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18~
## $ driverId        <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,~
## $ constructorId   <int> 1, 2, 3, 4, 1, 3, 5, 6, 2, 7, 8, 4, 6, 9, 7, 10, 9, 11~
## $ number          <int> 22, 3, 7, 5, 23, 8, 14, 1, 4, 12, 18, 6, 2, 9, 11, 20,~
## $ grid            <int> 1, 5, 7, 11, 3, 13, 17, 15, 2, 18, 19, 20, 4, 8, 6, 22~
## $ position        <int> 1, 2, 3, 4, 5, 6, 7, 8, NA, NA, NA, NA, NA, NA, NA, NA~
## $ positionText    <chr> "1", "2", "3", "4", "5", "6", "7", "8", "R", "R", "R",~
## $ positionOrder   <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,~
## $ points          <dbl> 10, 8, 6, 5, 4, 3, 2, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ laps            <int> 58, 58, 58, 58, 58, 57, 55, 53, 47, 43, 32, 30, 29, 25~
## $ time            <chr> "34:50.6", "5.478", "8.163", "17.181", "18.014", "", "~
## $ milliseconds    <int> 5690616, 5696094, 5698779, 5707797, 5708630, NA, NA, N~
## $ fastestLap      <int> 39, 41, 41, 58, 43, 50, 22, 20, 15, 23, 24, 20, 23, 21~
## $ rank            <int> 2, 3, 5, 7, 1, 14, 12, 4, 9, 13, 15, 16, 6, 11, 10, 17~
## $ fastestLapTime  <chr> "01:27.5", "01:27.7", "01:28.1", "01:28.6", "01:27.4",~
## $ fastestLapSpeed <chr> "218.3", "217.586", "216.719", "215.464", "218.385", "~
## $ statusId        <int> 1, 1, 1, 1, 1, 11, 5, 5, 4, 3, 7, 8, 5, 4, 10, 9, 4, 4~
```

```
##2.
results_tib1 <- results_tib %>% rename_with(tolower) %>% glimpse()
```

```
## Rows: 23,777
## Columns: 18
## $ resultid        <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,~
## $ raceid          <int> 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18~
## $ driverid        <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,~
## $ constructorid   <int> 1, 2, 3, 4, 1, 3, 5, 6, 2, 7, 8, 4, 6, 9, 7, 10, 9, 11~
## $ number          <int> 22, 3, 7, 5, 23, 8, 14, 1, 4, 12, 18, 6, 2, 9, 11, 20,~
## $ grid            <int> 1, 5, 7, 11, 3, 13, 17, 15, 2, 18, 19, 20, 4, 8, 6, 22~
## $ position        <int> 1, 2, 3, 4, 5, 6, 7, 8, NA, NA, NA, NA, NA, NA, NA, NA~
## $ positiontext    <chr> "1", "2", "3", "4", "5", "6", "7", "8", "R", "R", "R",~
## $ positionorder   <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,~
## $ points          <dbl> 10, 8, 6, 5, 4, 3, 2, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ laps            <int> 58, 58, 58, 58, 58, 57, 55, 53, 47, 43, 32, 30, 29, 25~
## $ time            <chr> "34:50.6", "5.478", "8.163", "17.181", "18.014", "", "~
## $ milliseconds    <int> 5690616, 5696094, 5698779, 5707797, 5708630, NA, NA, N~
## $ fastestlap      <int> 39, 41, 41, 58, 43, 50, 22, 20, 15, 23, 24, 20, 23, 21~
## $ rank            <int> 2, 3, 5, 7, 1, 14, 12, 4, 9, 13, 15, 16, 6, 11, 10, 17~
## $ fastestlaptime  <chr> "01:27.5", "01:27.7", "01:28.1", "01:28.6", "01:27.4",~
```

```
## $ fastestlapspeed <chr> "218.3", "217.586", "216.719", "215.464", "218.385", "~
## $ statusid        <int> 1, 1, 1, 1, 1, 11, 5, 5, 4, 3, 7, 8, 5, 4, 10, 9, 4, 4~
```

```
##3.
results_tib1 %>% count (resultid, raceid, driverid) %>% filter(n > 1)
```

```
## [1] resultid raceid   driverid n
## <0 rows> (or 0-length row.names)
```

- The uid's are resultid,raceid and driverid

```
##4.
results_tib2 <- results_tib1 %>% select(resultid , raceid, driverid, grid, position)
results_tib2 %>% glimpse()
```

```
## Rows: 23,777
## Columns: 5
## $ resultid <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18~
## $ raceid   <int> 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 1~
## $ driverid <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18~
## $ grid     <int> 1, 5, 7, 11, 3, 13, 17, 15, 2, 18, 19, 20, 4, 8, 6, 22, 14, 1~
## $ position <int> 1, 2, 3, 4, 5, 6, 7, 8, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
```

Iteration 2

Phase 1

1. Check for NA values in the columns
    - Drop NA
2. Mutate a new column for true or false values for same starting and final positions for qualifying and race leaders. 3.Check if tibble is Tidy

Phase 2

```
##1.
results_tib2 %>% summary()
```

```
##     resultid         raceid         driverid          grid
##  Min.   :    1   Min.   :  1.0   Min.   :  1.0   Min.   : 0.00
##  1st Qu.: 5945   1st Qu.:273.0   1st Qu.: 55.0   1st Qu.: 5.00
##  Median :11889   Median :478.0   Median :154.0   Median :11.00
##  Mean   :11889   Mean   :487.2   Mean   :226.5   Mean   :11.27
##  3rd Qu.:17833   3rd Qu.:718.0   3rd Qu.:314.0   3rd Qu.:17.00
##  Max.   :23781   Max.   :988.0   Max.   :843.0   Max.   :34.00
##
##     position
##  Min.   : 1.000
##  1st Qu.: 4.000
##  Median : 7.000
##  Mean   : 7.782
##  3rd Qu.:11.000
##  Max.   :33.000
##  NA's   :10550
```

```
results_tib3 <- results_tib2 %>% drop_na()
results_tib3 %>% summary()
```

```
##     resultid         raceid         driverid          grid
##  Min.   :    1   Min.   :  1.0   Min.   :  1.0   Min.   : 0.00
##  1st Qu.: 5254   1st Qu.:238.0   1st Qu.: 30.0   1st Qu.: 5.00
##  Median :12702   Median :494.0   Median :133.0   Median :11.00
##  Mean   :12280   Mean   :496.6   Mean   :229.9   Mean   :11.19
##  3rd Qu.:19399   3rd Qu.:772.0   3rd Qu.:328.0   3rd Qu.:17.00
##  Max.   :23779   Max.   :988.0   Max.   :843.0   Max.   :33.00
##     position
##  Min.   : 1.000
##  1st Qu.: 4.000
##  Median : 7.000
##  Mean   : 7.782
##  3rd Qu.:11.000
##  Max.   :33.000
```

```
##2.
results_tib4 <- results_tib3 %>% mutate(pos = if_else((grid == 1 & position == 1 ),1 ,0))
results_tib4 %>% glimpse()
```

```
## Rows: 13,227
## Columns: 6
## $ resultid <int> 1, 2, 3, 4, 5, 6, 7, 8, 23, 24, 25, 26, 27, 28, 29, 30, 31, 3~
## $ raceid   <int> 18, 18, 18, 18, 18, 18, 18, 18, 19, 19, 19, 19, 19, 19, 19, 1~
## $ driverid <int> 1, 2, 3, 4, 5, 6, 7, 8, 8, 9, 5, 15, 1, 2, 17, 4, 14, 18, 12,~
## $ grid     <int> 1, 5, 7, 11, 3, 13, 17, 15, 2, 4, 8, 3, 9, 5, 6, 7, 12, 11, 1~
## $ position <int> 1, 2, 3, 4, 5, 6, 7, 8, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12~
## $ pos      <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
```

```
##3.
results_tib4 %>% head(10)
```

```
##    resultid raceid driverid grid position pos
## 1         1     18        1    1        1   1
## 2         2     18        2    5        2   0
## 3         3     18        3    7        3   0
## 4         4     18        4   11        4   0
## 5         5     18        5    3        5   0
## 6         6     18        6   13        6   0
## 7         7     18        7   17        7   0
## 8         8     18        8   15        8   0
## 9        23     19        8    2        1   0
## 10       24     19        9    4        2   0
```

Race Information

Iteration 1

Phase 1

1. Read the csv file into R

2. Make column names lowercase
3. Identify uids
4. Drop unnecessary columns

Phase 2

```
##1.
races_tib <- read.csv("./Data/races.csv") %>% glimpse()
```

```
## Rows: 997
## Columns: 8
## $ raceId    <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 1~
## $ year      <int> 2009, 2009, 2009, 2009, 2009, 2009, 2009, 2009, 2009, 2009, ~
## $ round     <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 1~
## $ circuitId <int> 1, 2, 17, 3, 4, 6, 5, 9, 20, 11, 12, 13, 14, 15, 22, 18, 24,~
## $ name      <chr> "Australian Grand Prix", "Malaysian Grand Prix", "Chinese Gr~
## $ date      <chr> "2009-03-29", "2009-04-05", "2009-04-19", "2009-04-26", "200~
## $ time      <chr> "06:00:00", "09:00:00", "07:00:00", "12:00:00", "12:00:00", ~
## $ url       <chr> "http://en.wikipedia.org/wiki/2009_Australian_Grand_Prix", "~
```

```
##2.
races_tib1 <- races_tib %>% rename_with(tolower) %>% glimpse()
```

```
## Rows: 997
## Columns: 8
## $ raceid    <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 1~
## $ year      <int> 2009, 2009, 2009, 2009, 2009, 2009, 2009, 2009, 2009, 2009, ~
## $ round     <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 1~
## $ circuitid <int> 1, 2, 17, 3, 4, 6, 5, 9, 20, 11, 12, 13, 14, 15, 22, 18, 24,~
## $ name      <chr> "Australian Grand Prix", "Malaysian Grand Prix", "Chinese Gr~
## $ date      <chr> "2009-03-29", "2009-04-05", "2009-04-19", "2009-04-26", "200~
## $ time      <chr> "06:00:00", "09:00:00", "07:00:00", "12:00:00", "12:00:00", ~
## $ url       <chr> "http://en.wikipedia.org/wiki/2009_Australian_Grand_Prix", "~
```

```
##3.
races_tib1 %>% count (raceid, year) %>% filter(n > 1)
```

```
## [1] raceid year   n
## <0 rows> (or 0-length row.names)
```

- The uids are raceid and year

```
##4.
races_tib2 <- races_tib1 %>% select(-round, -time, -url)
races_tib2 %>% glimpse()
```

```
## Rows: 997
## Columns: 5
## $ raceid    <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 1~
## $ year      <int> 2009, 2009, 2009, 2009, 2009, 2009, 2009, 2009, 2009, 2009, ~
## $ circuitid <int> 1, 2, 17, 3, 4, 6, 5, 9, 20, 11, 12, 13, 14, 15, 22, 18, 24,~
## $ name      <chr> "Australian Grand Prix", "Malaysian Grand Prix", "Chinese Gr~
## $ date      <chr> "2009-03-29", "2009-04-05", "2009-04-19", "2009-04-26", "200~
```

Iteration 2

Phase 1

1. Check for NA values in the columns
    - Drop NA
2. Convert date column from character to Date & year column from int to Date

```
##1.
races_tib2 %>% summary()
```

```
##      raceid          year        circuitid         name
##  Min.   :   1   Min.   :1950   Min.   : 1.00   Length:997
##  1st Qu.: 250   1st Qu.:1974   1st Qu.: 9.00   Class :character
##  Median : 499   Median :1990   Median :18.00   Mode  :character
##  Mean   : 500   Mean   :1989   Mean   :21.76
##  3rd Qu.: 748   3rd Qu.:2005   3rd Qu.:30.00
##  Max.   :1009   Max.   :2018   Max.   :73.00
##      date
##  Length:997
##  Class :character
##  Mode  :character
##
##
##
```

```
##2.
races_tib3 <- races_tib2
races_tib3$date <- as.Date(races_tib3$date, "%Y-%m-%d")
races_tib3$year <-  ymd(races_tib3$year,truncated = 2L)
races_tib3 %>% glimpse()
```

```
## Rows: 997
## Columns: 5
## $ raceid    <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 1~
## $ year      <date> 2009-01-01, 2009-01-01, 2009-01-01, 2009-01-01, 2009-01-01,~
## $ circuitid <int> 1, 2, 17, 3, 4, 6, 5, 9, 20, 11, 12, 13, 14, 15, 22, 18, 24,~
## $ name      <chr> "Australian Grand Prix", "Malaysian Grand Prix", "Chinese Gr~
## $ date      <date> 2009-03-29, 2009-04-05, 2009-04-19, 2009-04-26, 2009-05-10,~
```

Iteration 3

Phase 1

1. Left join results and races tibble by uid (raceid)

Phase 2

```
join1 <- left_join(results_tib4, races_tib3, by="raceid")
join1 %>% glimpse()
```

```
## Rows: 13,227
## Columns: 10
## $ resultid  <int> 1, 2, 3, 4, 5, 6, 7, 8, 23, 24, 25, 26, 27, 28, 29, 30, 31, ~
## $ raceid    <int> 18, 18, 18, 18, 18, 18, 18, 18, 19, 19, 19, 19, 19, 19, 19, ~
## $ driverid  <int> 1, 2, 3, 4, 5, 6, 7, 8, 8, 9, 5, 15, 1, 2, 17, 4, 14, 18, 12~
## $ grid      <int> 1, 5, 7, 11, 3, 13, 17, 15, 2, 4, 8, 3, 9, 5, 6, 7, 12, 11, ~
## $ position  <int> 1, 2, 3, 4, 5, 6, 7, 8, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 1~
## $ pos       <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ year      <date> 2008-01-01, 2008-01-01, 2008-01-01, 2008-01-01, 2008-01-01,~
## $ circuitid <int> 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ~
## $ name      <chr> "Australian Grand Prix", "Australian Grand Prix", "Australia~
## $ date      <date> 2008-03-16, 2008-03-16, 2008-03-16, 2008-03-16, 2008-03-16,~
```

Safety Cars Deployed Data

Iteration 1

Phase 1

1. Read the csv file into R
2. Make column names lowercase
3. Determine if the data is Tidy and if not fix it
4. Identify uids
5. Drop unnecessary columns

Phase 2

```
##1.
sf_tib <- read.csv("./Data/safety_cars.csv") %>% glimpse()
```

```
## Rows: 312
## Columns: 5
## $ Race      <chr> "1973 Canadian Grand Prix", "1993 Brazilian Grand Prix", "19~
## $ Cause     <chr> "Accident", "Accident/Rain", "Stranded car", "Accident", "Ra~
## $ Deployed  <int> 33, 29, 38, 1, 28, 28, 13, 1, 1, 1, 15, 20, 42, 1, 28, 15, 2~
## $ Retreated <int> 39, 38, 40, 6, 33, 33, 18, 6, 4, 6, 18, 22, 49, 4, 33, 17, 2~
## $ FullLaps  <int> 5, 8, 1, 4, 4, 4, 4, 4, 3, 4, 2, 1, 6, 2, 4, 1, 4, 1, 4, 5, ~
```

```
##2.
sf_tib1 <- sf_tib %>% rename_with(tolower) %>% glimpse()
```

```
## Rows: 312
## Columns: 5
## $ race      <chr> "1973 Canadian Grand Prix", "1993 Brazilian Grand Prix", "19~
## $ cause     <chr> "Accident", "Accident/Rain", "Stranded car", "Accident", "Ra~
## $ deployed  <int> 33, 29, 38, 1, 28, 28, 13, 1, 1, 1, 15, 20, 42, 1, 28, 15, 2~
## $ retreated <int> 39, 38, 40, 6, 33, 33, 18, 6, 4, 6, 18, 22, 49, 4, 33, 17, 2~
## $ fulllaps  <int> 5, 8, 1, 4, 4, 4, 4, 4, 3, 4, 2, 1, 6, 2, 4, 1, 4, 1, 4, 5, ~
```

```
##3.
sf_tib2 <- sf_tib1 %>% separate(race, c("year", "name"), sep ="\\s", extra = "merge")
sf_tib2 %>% glimpse()
```

```
## Rows: 312
## Columns: 6
## $ year      <chr> "1973", "1993", "1993", "1994", "1995", "1996", "1996", "199~
## $ name      <chr> "Canadian Grand Prix", "Brazilian Grand Prix", "British Gran~
## $ cause     <chr> "Accident", "Accident/Rain", "Stranded car", "Accident", "Ra~
## $ deployed  <int> 33, 29, 38, 1, 28, 28, 13, 1, 1, 1, 15, 20, 42, 1, 28, 15, 2~
## $ retreated <int> 39, 38, 40, 6, 33, 33, 18, 6, 4, 6, 18, 22, 49, 4, 33, 17, 2~
## $ fulllaps  <int> 5, 8, 1, 4, 4, 4, 4, 4, 3, 4, 2, 1, 6, 2, 4, 1, 4, 1, 4, 5, ~
```

- This data set is from a different source hence I am creatiing a foreign key to relate and join the other csv files.

```
##4.
sf_tib2 %>% count (name, cause) %>% filter(n > 1) %>% head(5)
```

```
##                    name        cause  n
## 1  Abu Dhabi Grand Prix     Accident  5
## 2  Argentine Grand Prix     Accident  2
## 3 Australian Grand Prix     Accident 18
## 4 Australian Grand Prix Stranded car  7
## 5   Austrian Grand Prix     Accident  3
```

- There are no uid's
- name is a foreign key to the races.csv file.

```
##5.
sf_tib3 <- sf_tib2 %>% select(-retreated, -fulllaps)
sf_tib3 %>% glimpse()
```

```
## Rows: 312
## Columns: 4
## $ year     <chr> "1973", "1993", "1993", "1994", "1995", "1996", "1996", "1997~
## $ name     <chr> "Canadian Grand Prix", "Brazilian Grand Prix", "British Grand~
## $ cause    <chr> "Accident", "Accident/Rain", "Stranded car", "Accident", "Rai~
## $ deployed <int> 33, 29, 38, 1, 28, 28, 13, 1, 1, 1, 15, 20, 42, 1, 28, 15, 20~
```

Iteration 2

Phase 1

1. Check for NA values in the columns
    - Drop NA
2. Convert year to Date class
3. Create a new tibble with number of times a safety car has been deployed in the race

Phase 2

```
##1.
sf_tib3 %>% summary()
```

```
##      year               name              cause              deployed
##  Length:312         Length:312         Length:312         Min.   : 1.00
##  Class :character   Class :character   Class :character   1st Qu.: 1.00
##  Mode  :character   Mode  :character   Mode  :character   Median :15.50
##                                                           Mean   :19.92
##                                                           3rd Qu.:33.00
##                                                           Max.   :75.00
```

```
##2.
sf_tib4 <- sf_tib3 %>% mutate(year = ymd(year,truncated = 2L))
sf_tib4 %>% glimpse()
```

```
## Rows: 312
## Columns: 4
## $ year     <date> 1973-01-01, 1993-01-01, 1993-01-01, 1994-01-01, 1995-01-01, ~
## $ name     <chr> "Canadian Grand Prix", "Brazilian Grand Prix", "British Grand~
## $ cause    <chr> "Accident", "Accident/Rain", "Stranded car", "Accident", "Rai~
## $ deployed <int> 33, 29, 38, 1, 28, 28, 13, 1, 1, 1, 15, 20, 42, 1, 28, 15, 20~
```

```
##3.
sf_tib5 <- sf_tib3 %>% count(name)
sf_tib5 %<>% rename(count = n)
sf_tib5 %>% glimpse()
```

```
## Rows: 38
## Columns: 2
## $ name  <chr> "Abu Dhabi Grand Prix", "Argentine Grand Prix", "Australian Gran~
## $ count <int> 6, 2, 28, 12, 7, 8, 21, 22, 19, 28, 10, 1, 1, 3, 4, 3, 10, 7, 10~
```

Iteration 3

Phase 1

1. Left join races tibble and circuits tibble by uid(circuitid)
2. Left join safety cars tibble to the resulting tibble by uid (name)
3. Left join modified safety cars tibble to the resulting tibble by uid (name, year)

Phase 2

```
##1.
join3 <- left_join(races_tib3, circuits_tib3, by="circuitid")
join3 %>% glimpse()
```

```
## Rows: 997
## Columns: 9
## $ raceid     <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, ~
## $ year       <date> 2009-01-01, 2009-01-01, 2009-01-01, 2009-01-01, 2009-01-01~
## $ circuitid  <int> 1, 2, 17, 3, 4, 6, 5, 9, 20, 11, 12, 13, 14, 15, 22, 18, 24~
## $ name       <chr> "Australian Grand Prix", "Malaysian Grand Prix", "Chinese G~
## $ date       <date> 2009-03-29, 2009-04-05, 2009-04-19, 2009-04-26, 2009-05-10~
## $ cname      <chr> "Albert Park Grand Prix Circuit", "Sepang International Cir~
## $ circuitref <chr> "albert_park", "sepang", "shanghai", "bahrain", "catalunya"~
## $ location   <chr> "Melbourne", "Kuala Lumpur", "Shanghai", "Sakhir", "Montmel~
## $ country    <chr> "Australia", "Malaysia", "China", "Bahrain", "Spain", "Mona~
```

```r
##2.
join4 <- left_join(join3, sf_tib5, by=c("name"))
join4 %>% glimpse()
```

```
## Rows: 997
## Columns: 10
## $ raceid     <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, ~
## $ year       <date> 2009-01-01, 2009-01-01, 2009-01-01, 2009-01-01, 2009-01-01~
## $ circuitid  <int> 1, 2, 17, 3, 4, 6, 5, 9, 20, 11, 12, 13, 14, 15, 22, 18, 24~
## $ name       <chr> "Australian Grand Prix", "Malaysian Grand Prix", "Chinese G~
## $ date       <date> 2009-03-29, 2009-04-05, 2009-04-19, 2009-04-26, 2009-05-10~
## $ cname      <chr> "Albert Park Grand Prix Circuit", "Sepang International Cir~
## $ circuitref <chr> "albert_park", "sepang", "shanghai", "bahrain", "catalunya"~
## $ location   <chr> "Melbourne", "Kuala Lumpur", "Shanghai", "Sakhir", "Montmel~
## $ country    <chr> "Australia", "Malaysia", "China", "Bahrain", "Spain", "Mona~
## $ count      <int> 28, 6, 10, 8, 6, 23, 2, 19, 10, 7, 4, 21, 10, 20, 13, 22, 6~
```

```r
##3.
join5 <- left_join(join3, sf_tib4, by=c("name","year"))
join5 %>% glimpse()
```

```
## Rows: 1,069
## Columns: 11
## $ raceid     <int> 1, 1, 2, 3, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16~
## $ year       <date> 2009-01-01, 2009-01-01, 2009-01-01, 2009-01-01, 2009-01-01~
## $ circuitid  <int> 1, 1, 2, 17, 17, 3, 4, 6, 5, 9, 20, 11, 12, 13, 14, 15, 22,~
## $ name       <chr> "Australian Grand Prix", "Australian Grand Prix", "Malaysia~
## $ date       <date> 2009-03-29, 2009-03-29, 2009-04-05, 2009-04-19, 2009-04-19~
## $ cname      <chr> "Albert Park Grand Prix Circuit", "Albert Park Grand Prix C~
## $ circuitref <chr> "albert_park", "albert_park", "sepang", "shanghai", "shangh~
## $ location   <chr> "Melbourne", "Melbourne", "Kuala Lumpur", "Shanghai", "Shan~
## $ country    <chr> "Australia", "Australia", "Malaysia", "China", "China", "Ba~
## $ cause      <chr> "Accident", "Accident", "Rain", "Rain", "Debris from accide~
## $ deployed   <int> 19, 55, 31, 1, 18, NA, 1, NA, NA, NA, NA, NA, NA, 1, 53, 21~
```

```r
##4.
join6 <- left_join(races_tib3, sf_tib4, by=c("name","year"))
join6 %<>% select(-circuitid, -cause, -date) %>% drop_na()
join6 %>% glimpse()
```

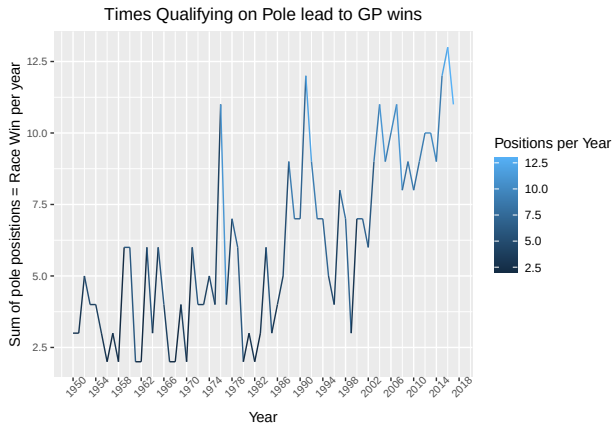
```
## Rows: 237
## Columns: 4
## $ raceid   <int> 1, 1, 2, 3, 3, 5, 12, 13, 14, 15, 16, 18, 18, 18, 21, 21, 22,~
## $ year     <date> 2009-01-01, 2009-01-01, 2009-01-01, 2009-01-01, 2009-01-01, ~
## $ name     <chr> "Australian Grand Prix", "Australian Grand Prix", "Malaysian ~
## $ deployed <int> 19, 55, 31, 1, 18, 1, 1, 53, 21, 45, 1, 1, 26, 44, 1, 22, 1, ~
```

3. Results/Discussion

Question 1:

```
p1 <- join1 %>% group_by(year) %>% summarize(posperyear = sum(pos)) %>% ungroup() %>% ggplot() +
  geom_line(mapping = aes (x = year, y = posperyear, color = posperyear )) +
  theme_gray() +
  labs(title = "Times Qualifying on Pole lead to GP wins",
       y = "Sum of pole posistions = Race Win per year",
       x = "Year",
       color = "Positions per Year") +
  scale_x_date(breaks="4 year", date_labels = "%Y") +
  scale_y_continuous(breaks = c(2.5, 5.0, 7.5, 10.0,12.5)) +
  theme(axis.text.x = element_text(size=8,angle=45),
        axis.text.y = element_text(size=8),
        plot.title = element_text(hjust = 0.5))

p1
```



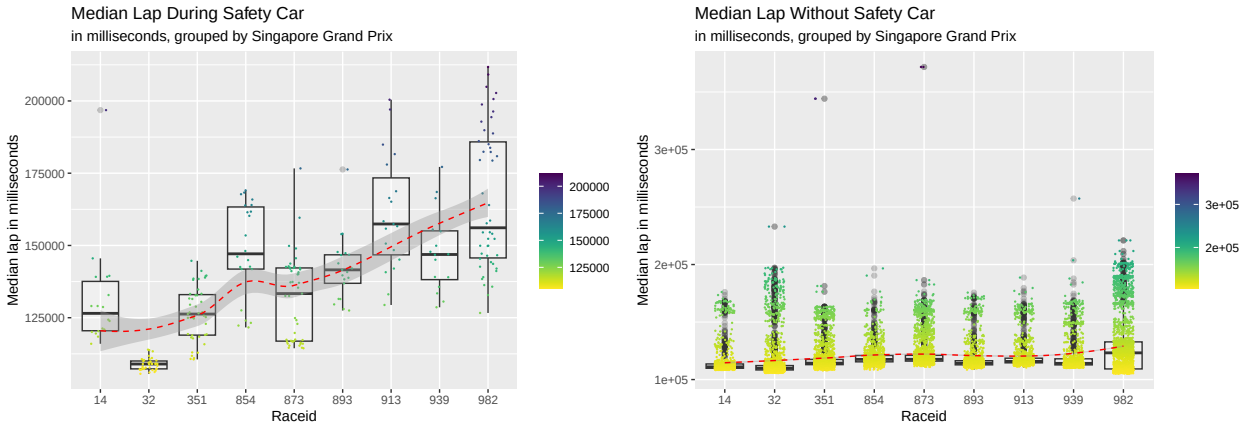
There is a general upward trend of drivers starting in pole position in qualifying sessions and finishing out on top in the Grand Prix on Sunday. The drops in the graph before the 20th century can be attributed to the existing technology of those days. The introduction of hybrid engines in 2014 has shown its effect as there is a significant upward trend in pole positions resulting in wins in the grand prix. Around 42.41% of race wins have been won by drivers on pole. The most amount of wins from pole positions is seen after 2014, after which it drops.

Question 2:

```
p2 <- join7 %>% filter(deployed ==lap, name =="Singapore Grand Prix") %>%
  group_by(raceid,deployed) %>%
  ggplot(aes(x = as.factor(raceid),y = milliseconds, color = milliseconds)) +
  geom_boxplot(alpha=.25) + theme_gray() +
  labs(title = "Median Lap During Safety Car",
       subtitle ='in milliseconds, grouped by Singapore Grand Prix',
       y = "Median lap in milliseconds",
       x = "Raceid",
       color = "Time") +
  geom_jitter(shape=16,position=position_jitter(0.2),size=0.5) +
  geom_smooth(method='loess',aes(group=1),color='red',lty=2,size=.5) +
  scale_color_gradientn(name="",colours=rev(viridis::viridis(20)))
```

```
p3 <- join7 %>% filter(deployed!=lap, name=="Singapore Grand Prix") %>%
  group_by(raceid,deployed) %>%
```

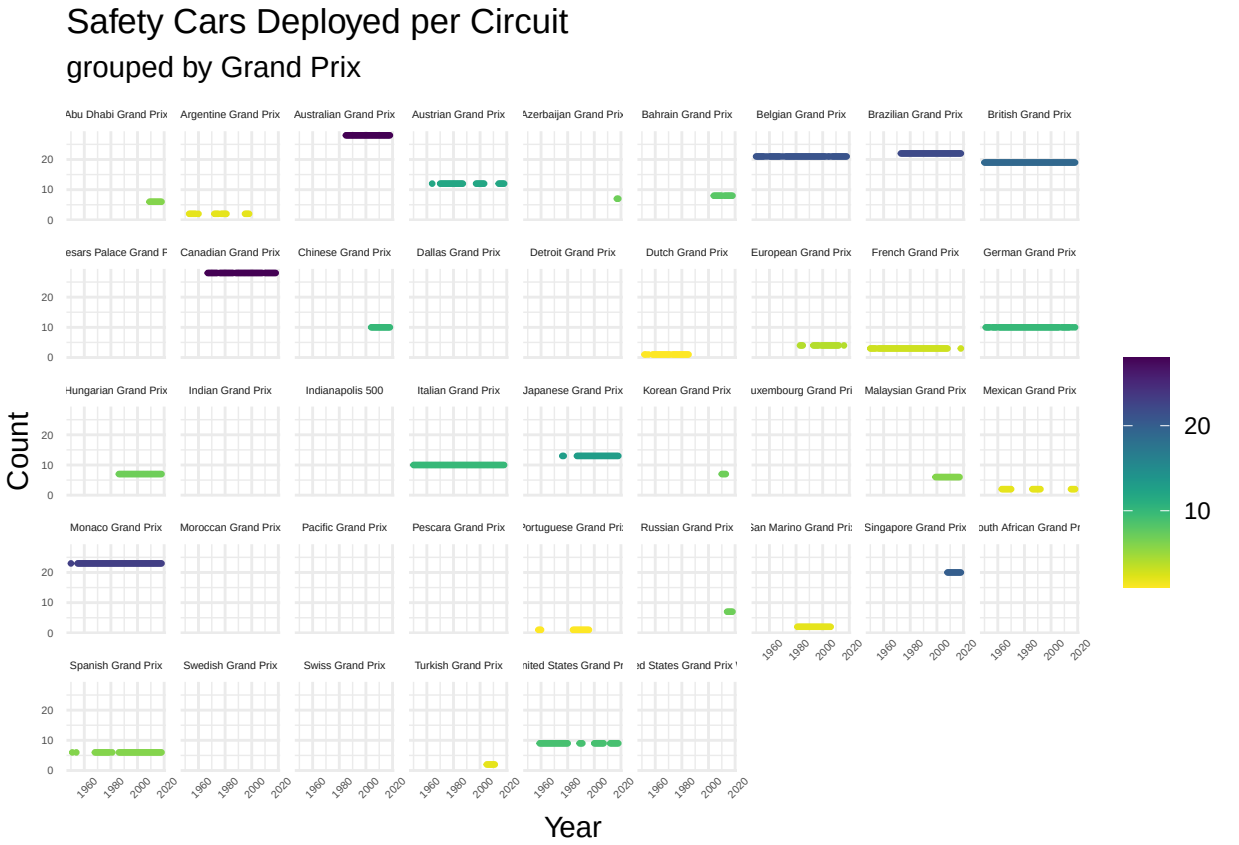
```
  ggplot(aes(x = as.factor(raceid),y = milliseconds, color = milliseconds)) +
  geom_boxplot(alpha=.25) + theme_gray() +
  labs(title = "Median Lap Without Safety Car",
       subtitle ='in milliseconds, grouped by Singapore Grand Prix',
       y = "Median lap in milliseconds",
       x = "Raceid",
       color = "Time") +
  geom_jitter(shape=16,position=position_jitter(0.2),size=0.5) +
  geom_smooth(method='loess',aes(group=1),color='red',lty=2,size=.5) +
  scale_color_gradientn(name="",colours=rev(viridis::viridis(20)))
```



I chose to analyse the Singapore races as this is the circuit with the most number of safety car deployments. Deployment of safety cars during the race causes the cars to slow down and take more time to complete a lap as they have to maintain a speed. This causes the cars to bunch up behind the safety car, although they have to maintain a minimum car length in between. Without the safety car, the cars would lap the Singapore circuit in under a minute.

Question 3:

```
p4 <- join4 %>% group_by(name,year) %>%
  ggplot(aes(x=year,y= count, color=count)) +
  geom_point(size=0.5) + theme_minimal() +
  labs(title = "Safety Cars Deployed per Circuit",
       subtitle = "grouped by Grand Prix",
       y = "Count",
       x = "Year",
       color = "Count") +
  scale_color_gradientn(name="",colours=rev(viridis::viridis(20))) +
  theme(axis.text.x = element_text(size=4,angle=45),
        axis.text.y = element_text(size=4),
        strip.text.x = element_text(size = 4)) + facet_wrap(~name,ncol=9)
p4
```



Clearly, there are certain circuits with Safety Cars deployed frequently. The Belgian Grand Prix, British Grand Prix, German Grand Prix, Monaco Grand Prix, and Italian Grand Prix are the circuits with the most number of safety cars deployed. From my observations above, the Singapore Grand Prix has deployed a safety car in its circuit every year a race is held there.

Conclusion

1. Statistics show that the pole sitter takes more than 40% of race wins. The ability to start on pole position can be extremely advantageous for F1 drivers as it gives the driver a shorter run to the first corner. The pole position driver can simply focus on starting the race and would not be too concerned about the other drivers. As there is less traffic in-front of the leading driver they are more safe than the mid-fielders who are trying the attain a good position on track, apex and the score board. They are less in danger of being in an accident and deploying safety cars. As F1 progresses throughout the years with better technology and machinery the drivers on pole have more chance of wining the race safetly.

2. When a safety car is deployed the position of the cars on track does not change, but the time gap eliminates with the car behind. As overtaking is prohibited when a safety car is on track, overtaking is much easier when the race resumes. This also gives time for drivers to pit for new fresh tyres and give a tire advantage when the race resumes. Since the cars are forced to maintain a specific speed it is easier for th car at the back to catch up. Once the safety car is removed, the race will be close. Hence, the safety car is advantageous to the cars at the back but disadvantageous to those leading the race.

3. In 2023, the Italian Grand Prix, Monaco Grand Prix, British Grand Prix, Belgian Grand Prix, and Singapore Grand Prix are still upcoming. Surprisingly the circuits with the most number of safety cars deployed and red flags have been on the race calender almost every year.Some of the circuits are included in the calendar every year because these circuits with their high-speed corners and deep-narrow turns are challenging for teams to race in every year, which provides the thrill of the sport.

The analysis could be improved by analyzing the events every lap that lead to a safety car being deployed and examining other factors that would lead to a safety car appearing in a race.

References

1. Jason Hope (2022). How Often Does the Pole-Sitter Win In F1? F1Chronicle.com
2. Safety Car Deployed. Motorsports-regulations.com