Project

Brecon Welch

2023-12-11

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
f1 <- read_csv("archive (1)/status.csv")
```

```
## Rows: 139 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): status
## dbl (1): statusId
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
f1
```

```
## # A tibble: 139 x 2
##    statusId status
##       <dbl> <chr>
##  1        1 Finished
##  2        2 Disqualified
##  3        3 Accident
##  4        4 Collision
##  5        5 Engine
##  6        6 Gearbox
##  7        7 Transmission
##  8        8 Clutch
##  9        9 Hydraulics
## 10       10 Electrical
## # i 129 more rows
```

```
f1_drivers <- read_csv("formula1-dataset-master/formula1_data_races.csv")
```

```
## Rows: 23175 Columns: 10
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (6): position, driver, car, time, grand_prix, fastest_lap
## dbl (4): number, laps, points, year
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
f1_drivers
```

```
## # A tibble: 23,175 x 10
##    position number driver  car    laps time  points grand_prix  year fastest_lap
##    <chr>     <dbl> <chr>   <chr> <dbl> <chr>  <dbl> <chr>      <dbl> <chr>
##  1 1            44 Lewis ~ Merc~    56 1:32~     25 BAHRAIN     2021 no
##  2 2            33 Max  V~ Red ~    56 +0.7~     18 BAHRAIN     2021 no
##  3 3            77 Valtte~ Merc~    56 +37.~     16 BAHRAIN     2021 yes
##  4 4             4 Lando ~ McLa~    56 +46.~     12 BAHRAIN     2021 no
##  5 5            11 Sergio~ Red ~    56 +52.~     10 BAHRAIN     2021 no
##  6 6            16 Charle~ Ferr~    56 +59.~      8 BAHRAIN     2021 no
##  7 7             3 Daniel~ McLa~    56 +66.~      6 BAHRAIN     2021 no
##  8 8            55 Carlos~ Ferr~    56 +67.~      4 BAHRAIN     2021 no
##  9 9            22 Yuki  ~ Alph~    56 +85.~      2 BAHRAIN     2021 no
## 10 10           18 Lance ~ Asto~    56 +86.~      1 BAHRAIN     2021 no
## # i 23,165 more rows
```

```
F1D <- read_csv("F1Drivers_Dataset.csv")
```

```
## Rows: 868 Columns: 22
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (4): Driver, Nationality, Seasons, Championship Years
## dbl (16): Championships, Race_Entries, Race_Starts, Pole_Positions, Race_Win...
## lgl  (2): Active, Champion
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
F1D
```

```
## # A tibble: 868 x 22
##    Driver            Nationality  Seasons Championships Race_Entries Race_Starts
##    <chr>             <chr>        <chr>           <dbl>        <dbl>       <dbl>
##  1 Carlo Abate       Italy        [1962,~             0            3           0
##  2 George Abecassis  United King~ [1951,~             0            2           2
##  3 Kenny Acheson     United King~ [1983,~             0           10           3
##  4 Andrea de Adamich Italy        [1968,~             0           36          30
##  5 Philippe Adams    Belgium      [1994]              0            2           2
##  6 Walt Ader         United Stat~ [1950]              0            1           1
```

```
##  7 Kurt Adolff       West Germany [1953]              0           1           1
##  8 Fred Agabashian   United Stat~ [1950,~             0           9           8
##  9 Kurt Ahrens Jr.   West Germany [1966,~             0           4           4
## 10 Jack Aitken       United King~ [2020]              0           1           1
## # i 858 more rows
## # i 16 more variables: Pole_Positions <dbl>, Race_Wins <dbl>, Podiums <dbl>,
## #   Fastest_Laps <dbl>, Points <dbl>, Active <lgl>, `Championship Years` <chr>,
## #   Decade <dbl>, Pole_Rate <dbl>, Start_Rate <dbl>, Win_Rate <dbl>,
## #   Podium_Rate <dbl>, FastLap_Rate <dbl>, Points_Per_Entry <dbl>,
## #   Years_Active <dbl>, Champion <lgl>
```

```
races <- read_csv("races.csv")
```

```
## Rows: 1101 Columns: 18
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (13): name, time, url, fp1_date, fp1_time, fp2_date, fp2_time, fp3_date...
## dbl   (4): raceId, year, round, circuitId
## date  (1): date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
drivers <- read_csv("drivers.csv")
```

```
## Rows: 857 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): driverRef, number, code, forename, surname, nationality, url
## dbl  (1): driverId
## date (1): dob
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

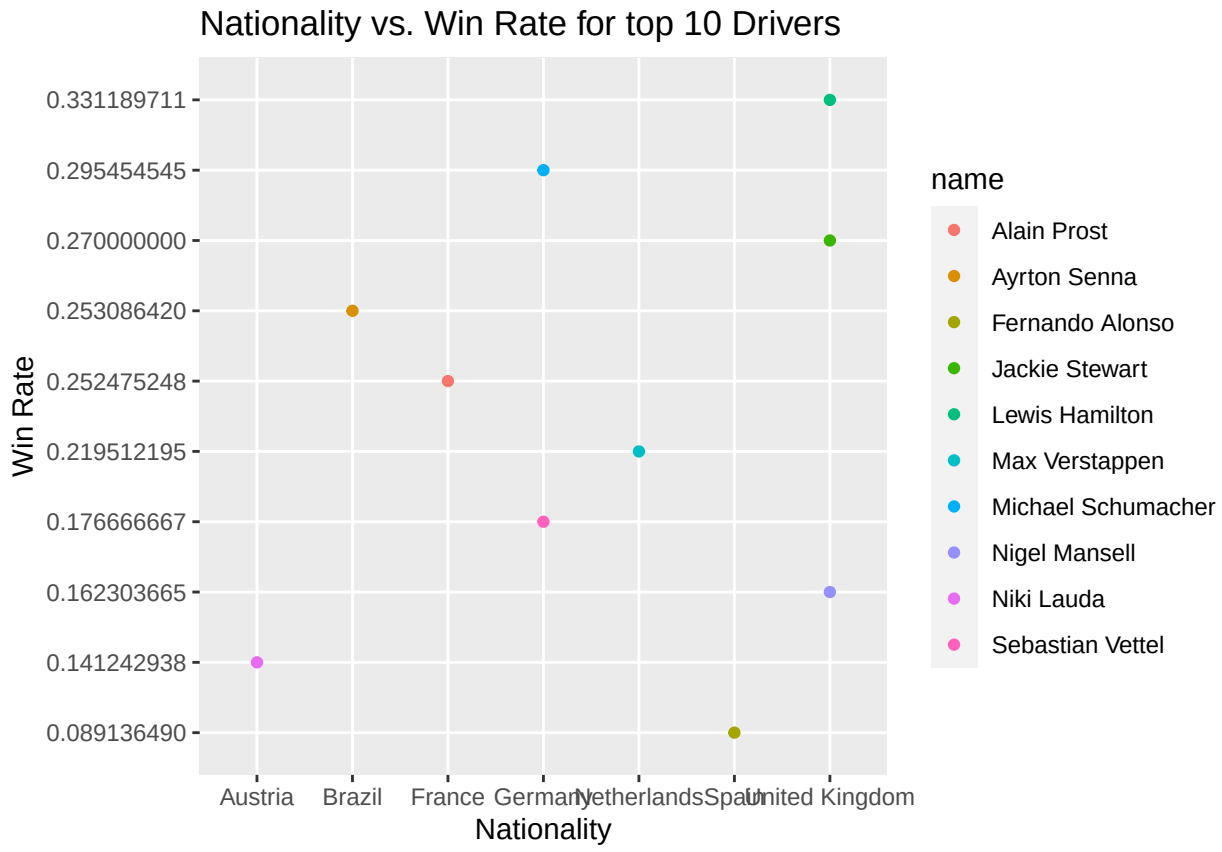
```
results <- read_csv("results.csv.zip")
```

```
## Warning: One or more parsing issues, call `problems()` on your data frame for details,
## e.g.:
##   dat <- vroom(...)
##   problems(dat)
```

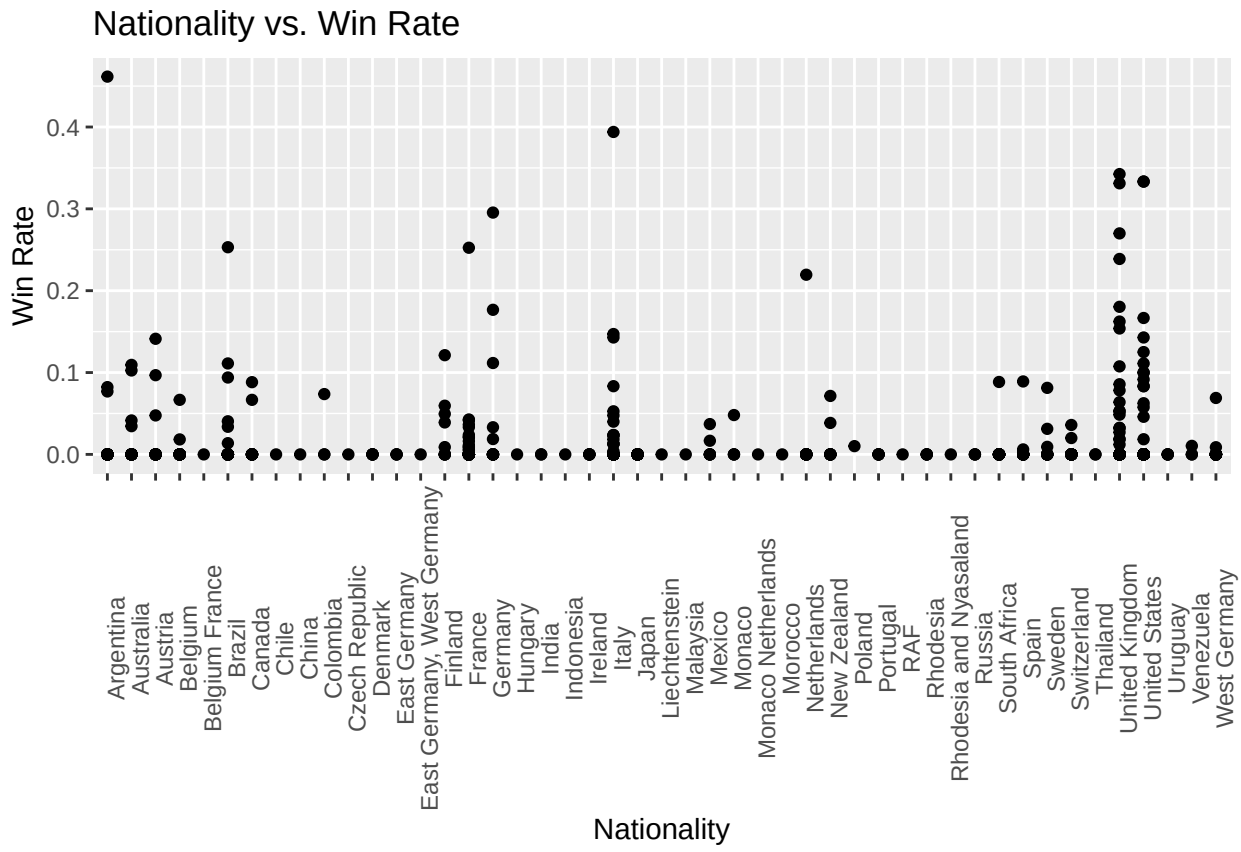
```
## Rows: 26080 Columns: 18
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (8): position, positionText, time, milliseconds, fastestLap, rank, fast...
## dbl (10): resultId, raceId, driverId, constructorId, number, grid, position0...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

##Plots

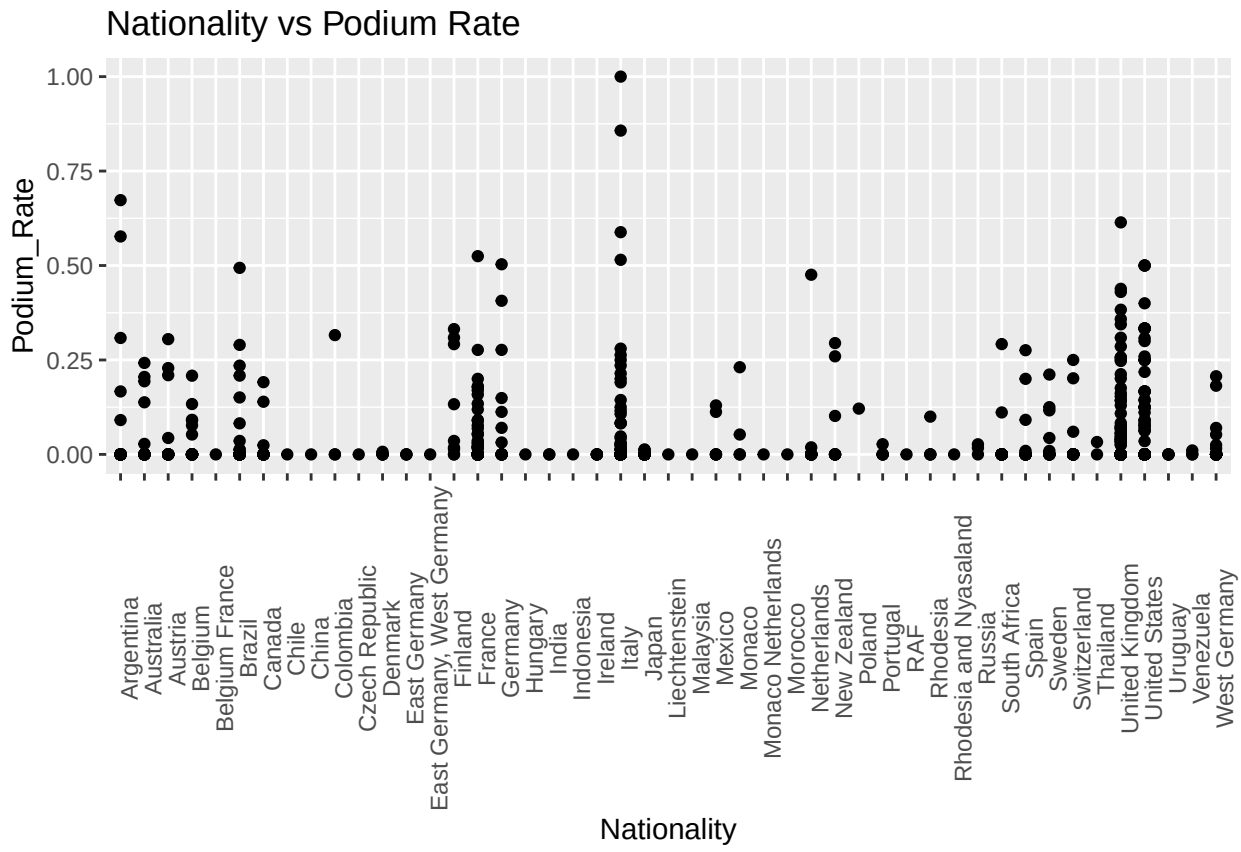
```
library(ggplot2)
F1data <- data.frame(
  name = c("Lewis Hamilton", "Michael Schumacher", "Sebastian Vettel", "Alain Prost", "Ayrton Senna", "
  Win_Rate = c("0.331189711", "0.295454545", "0.176666667", "0.252475248", "0.253086420", "0.219512195"
  Nationality = c("United Kingdom", "Germany", "Germany", "France", "Brazil", "Netherlands", "Spain", "

ggplot(F1data, aes(x = Nationality, y = Win_Rate, color = name)) +
  geom_point() +
  labs(title = "Nationality vs. Win Rate for top 10 Drivers", x = "Nationality", y = "Win Rate")
```



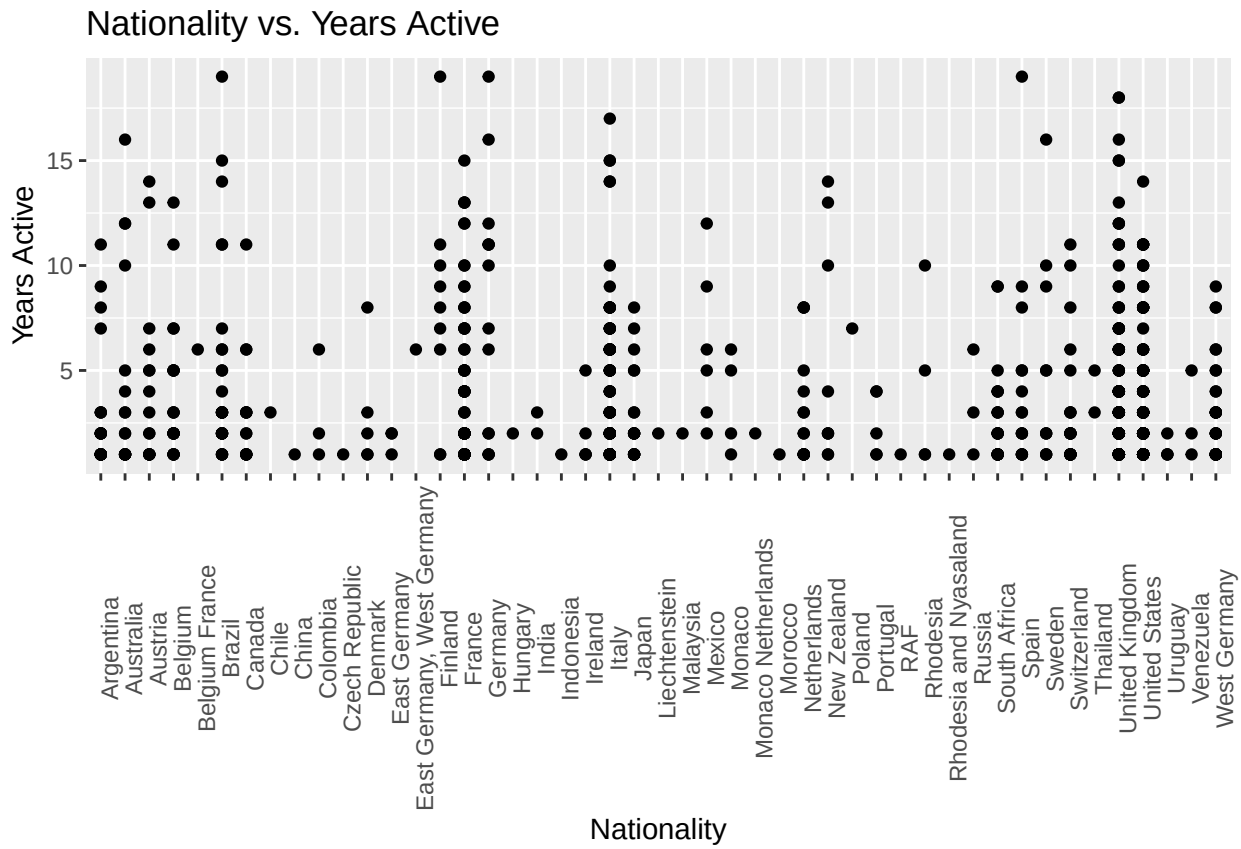

```
win <- ggplot(data = F1D,
       aes(x = Nationality, y = Win_Rate)) +
  geom_point() + labs(title = "Nationality vs. Win Rate", x = "Nationality", y = "Win Rate") + theme(ax
win
```




```
podium <- ggplot(data = F1D,
       aes(x = Nationality, y = Podium_Rate)) +
  geom_point() + labs(title = "Nationality vs Podium Rate", x = "Nationality", y = "Podium_Rate") + ther
podium
```




```
years <- ggplot(data = F1D,
       aes(x = Nationality, y = Years_Active)) +
  geom_point() + labs(title = "Nationality vs. Years Active", x = "Nationality", y = "Years Active") + t
years
```



##Regression

```
library(modelsummary)
```

```
## 'modelsummary' has built-in support to draw text-only (markdown) tables.
##   To generate tables in other formats, you must install one or more of
##   these libraries:
##
## install.packages(c(
##     "kableExtra",
##     "gt",
##     "flextable",
##
##   "huxtable",
##     "DT"
## ))
##
##   Alternatively, you can set markdown as the default table format to
##   silence this alert:
##
## config_modelsummary(factory_default = "markdown")
```

```
Fit <- lm(Win_Rate ~ Nationality + FastLap_Rate + Decade + Podium_Rate + Years_Active, data = F1D)

var_labels <- c(
```

```
"(Intercept)" = "Intercept",
"Win_Rate" = "Win Rate",
"Nationality" = "Nationality ",
"FastLap_Rate" = "Fasted Lap Rate",
"Decade" = "Decade of Racing",
"Podium_Rate" = "Podium Rate",
"Years_Active" = "Years Active"
)


modelsummary::modelsummary(Fit,
                           statistic = c("s.e. = {std.error}",
                                         "p = {p.value}"),
                           gof_map = c("nobs", "r.squared", "adj.r.squared"))
```

	(1)
(Intercept)	-0.017
	s.e. = 0.085
	p = 0.836
NationalityAustralia	0.004
	s.e. = 0.006
	p = 0.512
NationalityAustria	0.006
	s.e. = 0.007
	p = 0.322
NationalityBelgium	0.004
	s.e. = 0.006
	p = 0.504
NationalityBelgium France	-0.001
	s.e. = 0.021
	p = 0.952
NationalityBrazil	0.008
	s.e. = 0.006
	p = 0.139
NationalityCanada	0.007
	s.e. = 0.007
	p = 0.317
NationalityChile	0.007
	s.e. = 0.021
	p = 0.738
NationalityChina	-0.054
	s.e. = 0.021
	p = 0.009
NationalityColombia	-0.009
	s.e. = 0.012
	p = 0.450
NationalityCzech Republic	0.007
	s.e. = 0.021
	p = 0.748
NationalityDenmark	0.005
	s.e. = 0.010
	p = 0.637

	(1)
NationalityEast Germany	0.007
	s.e. = 0.012
	p = 0.574
NationalityEast Germany, West Germany	0.007
	s.e. = 0.021
	p = 0.726
NationalityFinland	-0.008
	s.e. = 0.008
	p = 0.326
NationalityFrance	0.004
	s.e. = 0.005
	p = 0.376
NationalityGermany	0.011
	s.e. = 0.007
	p = 0.124
NationalityHungary	0.007
	s.e. = 0.021
	p = 0.745
NationalityIndia	0.007
	s.e. = 0.015
	p = 0.655
NationalityIndonesia	0.007
	s.e. = 0.021
	p = 0.753
NationalityIreland	0.007
	s.e. = 0.010
	p = 0.491
NationalityItaly	0.005
	s.e. = 0.005
	p = 0.318
NationalityJapan	0.006
	s.e. = 0.006
	p = 0.347
NationalityLiechtenstein	0.007
	s.e. = 0.021
	p = 0.739
NationalityMalaysia	0.007
	s.e. = 0.021
	p = 0.745
NationalityMexico	0.003
	s.e. = 0.009
	p = 0.728
NationalityMonaco	0.000
	s.e. = 0.011
	p = 0.968
NationalityMonaco Netherlands	0.007
	s.e. = 0.021
	p = 0.748
NationalityMorocco	0.007
	s.e. = 0.021
	p = 0.740
NationalityNetherlands	0.012

	(1)
	s.e. = 0.007
	p = 0.074
NationalityNew Zealand	0.001
	s.e. = 0.008
	p = 0.915
NationalityPoland	-0.003
	s.e. = 0.021
	p = 0.887
NationalityPortugal	0.006
	s.e. = 0.010
	p = 0.529
NationalityRAF	0.007
	s.e. = 0.021
	p = 0.753
NationalityRhodesia	0.004
	s.e. = 0.011
	p = 0.687
NationalityRhodesia and Nyasaland	0.007
	s.e. = 0.021
	p = 0.740
NationalityRussia	-0.001
	s.e. = 0.013
	p = 0.940
NationalitySouth Africa	0.008
	s.e. = 0.006
	p = 0.198
NationalitySpain	0.005
	s.e. = 0.007
	p = 0.464
NationalitySweden	0.007
	s.e. = 0.007
	p = 0.374
NationalitySwitzerland	0.002
	s.e. = 0.006
	p = 0.684
NationalityThailand	0.005
	s.e. = 0.015
	p = 0.727
NationalityUnited Kingdom	0.007
	s.e. = 0.004
	p = 0.090
NationalityUnited States	0.005
	s.e. = 0.004
	p = 0.249
NationalityUruguay	0.007
	s.e. = 0.011
	p = 0.527
NationalityVenezuela	0.010
	s.e. = 0.012
	p = 0.422
NationalityWest Germany	0.005
	s.e. = 0.005

	(1)
	p = 0.382
FastLap_Rate	0.700
	s.e. = 0.022
	p = <0.001
Decade	0.000
	s.e. = 0.000
	p = 0.899
Podium_Rate	0.108
	s.e. = 0.009
	p = <0.001
Years_Active	0.000
	s.e. = 0.000
	p = 0.749
Num.Obs.	868
R2	0.797
R2 Adj.	0.784

My research question investigates whether nationality has a significant impact on Formula 1 drivers' win rate, controlling for variables such as Podium rate, fastest lap rate, years active, and the decade of competition. I hypothesize that even after accounting for these control factors, nationality may still influence how much a driver wins. This is because certain countries may have a stronger tradition, infrastructure, or support system for motorsports, contributing to the success of drivers from those nations. Furthermore, the availability of financial support and sponsorships for drivers, cultural attitudes toward risk-taking and competitiveness, and fan support vary per country, and thus have an impact. The explanatory variable is Nationality which is a categorical variable representing the country of origin of each Formula 1 driver. The outcome variable of interest is Win Rate. Win Rate is a continuous variable that measures the percentage of races a driver has won out of the total races they have participated in. My hypothesis would be supported if there is a statistically significant correlation between nationality and win rate with drivers from certain countries consistently exhibiting higher win rates compared to drivers from other countries. For instance, if drivers from a specific country consistently win a higher percentage of races, it suggests that nationality is a relevant factor in predicting win rates. However, if the difference in these variables is negligible, this would suggest that nationality alone may not be a significant predictor of Formula 1 drivers' win rates, and other factors might play a more influential role. Understanding this potential impact is crucial for unraveling the complexities that contribute to success in Formula 1. It can provide insights into whether certain nationalities have a consistent advantage or disadvantage, guiding teams in their strategic decisions.

The dataset I chose encompasses information on win rate, podium rate, fastest lap rate, decade, years active, and nationality of Formula 1 drivers. Podium Rate measures the percentage of races a driver has placed first, second, or third in out of the total races they have participated in. Fastest Lap Rate is defined as the frequency at which a driver achieves the fastest lap during a race or a specified time period. Decade refers to the decade in which the driver participated in races. The Years Active variable shows how long a driver has been racing in Formula 1. A plot visualizing the main outcome of interest, the distribution of win rates across different nationalities, offers an initial exploration. This plot serves as a visual guide to understanding potential variations in win rates among drivers from different countries.

According to the plot, three out of the top ten most winning Formula 1 drivers originate from the United Kingdom and two out of the top ten are from Germany. With two countries making up the nationality of half of the most successful drivers, it is rational to continue to explore the relationship between nationality and win rate to see if there is indeed a correlation between the two variables. However, this plot is limiting in that it only visualizes ten drivers. On one hand, looking at a larger pool of drivers could show more correlation between win rate and nationality if the same nationalities show up repeataedly at the top of the ranks. On the other hand, however, more data may show that this finding for the top 10 winningest drivers was merely a coincidence and that many of the top drivers are from different countries.

With further analysis comparing nationality with years active, podium rate, and win rate for all the drivers in the dataset, we can see that there are certain countries that win more often, place on the podium more freqeuently, and have longer careers. For podium and win rates these countries included, the United States, United Kingdom, Italy, France, and Germany. These countries in addition to Australia, Austria, Brazil, and West Germany were shown to have more drivers with longer careers compared to the other included countries.

While this analysis provides helpful insight, the main analysis involves a regression model examining the relationship between win rate and nationality, controlling for podium rate, fastest lap time, fastest lap speed, and decade. The regression output and accompanying plot illustrate the significance and direction of the relationship between nationality and win rate. Interpretation of the main coefficient of interest assesses whether nationality has a statistically significant impact on win rate after considering the control variables. Moreover, the coefficients associated with each level of the nationality variable indicate how much the win rate is expected to change for drivers from those countries compared to the reference category, while controlling for the other variables in the model. After running the regression, we get an average regression cofficient of 0.0035 for the 46 different nationalities included in the dataset.

The coefficient for nationality represents the estimated change in win rate associated with being from a specific country compared to the reference category (Argentina) while keeping the other variables constant. This means that holding all other variables constant, the estimated change in win rate for a one-unit change in Nationality from 0 to 1 is 0.0035, on average. From the data, we can see that the Netherlands had the highest coefficient (0.012). Since the reference category is Argentina and the coefficient for the Netherlands in the regression model is 0.012, this means that drivers from the Netherlands, on average, have a Win Rate that is 0.012 higher than drivers from Argentina, holding other variables constant. So, while from this data it seems that drivers from the Netherlands tend to have a slightly higher Win Rate compared to drivers from Argentina, it is important to look at the p-value to determine whether these results are stastitically significant. Since $p = 0.748$ and is greater than the siginificance level 0.05, this suggests that there is no strong evidence the nationality of the driver has a significant impact on their win rate. Thus, I believe that this coefficient does not represent a causal effect.

The regression analysis aimed to explore the impact of nationality on Formula 1 drivers' win rates while controlling for other relevant factors. The results reveal that the coefficient for nationality is not statistically significant ($p > 0.05$), suggesting that, within the context of this model and dataset, there is no strong evidence to support the hypothesis that nationality has a significant impact on win rates. The lack of statistical significance implies that, after accounting for variables like Fast Lap Rate, Decade, Podium Rate, and Years Active, Nationality alone does not appear to be a significant predictor of win rates for Formula 1 drivers.

A possible limitation could be that the analysis's generalizability might be limited if the sample does not adequately represent the entire population of Formula 1 drivers. While this dataset provides an extensive list of drivers and nationalities, more races have taken place and more drivers have begun racing in Formula 1 since the creation of the dataset. All of this new data could have an impact on future analysis of nationality and win rate, establishing a relationship that is not shown in the present analysis. There are also many confounding variables that have not been accounted for such as team dynamics, car performance, or individual skill levels that can influence a driver's success rate. Looking at whether certain nationalities tend to drive with certain teams could also provide more insight into the effects of nationality on win rate, and possibly explain some of the aforementioned confounding variables such as team dynamics.

To improve this analysis, I would collect more data so that my data set is up to date with the current racers. Having a larger data set would enhance the study's power, providing more reliable insights into the relationship between nationality and win rates. Along this same trend of a larger data set, including more relevant variables, such as age, team-related factors, car technology, and driver experience, could contribute to a more comprehensive analysis. Additionally, I think it could be interesting to use a Differences-in-Differences (DiD) design rather than a cross-sectional one like in the present study. DiD involves comparing changes in outcomes over time between a treatment group and a control group. In the context of F1, I could identify a group of drivers as the treatment group who experience a change in nationality, such as switching countries, and compare their win rate changes with a control group of drivers who do not experience such

changes. This could help control for time-related factors and other unobserved variables that may affect win rates.

While my analysis did not find a significant relationship between nationality and win rates, it was an interesting question to explore. Further research with a more extensive dataset and a nuanced approach could provide a deeper understanding of the complex dynamics influencing Formula 1 drivers' success.