Assignment 2

Benjámin Salga

Introduction

Nowadays, no secret acing with cars at a higher level of motorsport comes with a huge amount of money. In our history, some Formula 1 drivers have been considered the highest-paid athletes in the world of sport. The world has changed and today's sportsmen from football and other sports earn more. However, F1 drivers still get a hefty reward after the races. When drivers post images on social media of their lavish homes, fast cars, and hedonistic lifestyle, are great proof of that. On the other side, these people work hard for their money. Driving on such a high level is very exhausting and needs special skills, but everybody deserves the reward? Some drivers earn 3-5 Million dollars others earn 10 times that. Are those drivers 10 times better?

This assignment focuses on uncovering the probability that F1 drivers are getting more rewarded for various explanatory variables: such as the finish position, different teams the driver race with, driver age, and their nationality. For the sake of the analysis, the kaggle F1 world championship dataset was used. It includes information from 1950 to 2021, however, this analysis only focuses on the Hybrid area and on drivers who raced in 2021. The hybrid era means the area between 2014 and 2021, where the teams are allowed to use only 1.6-liter hybrid v6 engines. This area has been ended whit the 2021 season, which is why I chose this topic for the analysis.

The information about the wages is collected by myself from Sporttrac year by year(https://www.spotrac.com/formula1/2020/). The earnings contain all the bonuses, but only from Formula 1. Payment from other companies is not included in the salary.

Data summary and cleaning

After the data has been loaded downloaded from kaggle (https://www.kaggle.com/rohanrao/formula-1-world-championship-1950-2020), the data was cleaned and merged, after that I checked the summary of the variables (e.g. Finish position at race, Finish position at qualification, Fastest lap times at race, Earned points by the driver, Earned points by teams) I intended to use in our analysis. The data summary showed the borders of the data set, which seems to match with reality. Grid, Qulay position and position order is between 1 and 22, which is valid since before 2016 there was a team called Marussia F1. They are the only team in the sport, who did not return to Formula 1 (Many other teams changed their name and owners, only Marussia did not). The Mean shows us that a bigger part of the variable is less than 10 since the analysis focus on drivers in 2021 and 4 World champion were in this season, which cause this change. The same rule is true for the point earned by the drivers' observation since its Mean shows the same pattern, however in this case the Max shows an extreme value since only once at the end of the season in Abu Dhabi, the FIA decided to give double point for the driver this is why instead of the regular 25 points the Max is 50 points.

There is an error in the dataset, which comes from bad recording probably: The 'Wins' variable, the number of first places by teams are not correct. It is proved on the chart (Appendix Chart 1). The provided information by the chart shows that Red Bull has the most wined races with 47 trophies and Mercedes AMG is only third around 27 races, which is interesting since only Hamilton had won more than 60 races during this period. This variable was not used in the analysis and there was no issue with other variables.

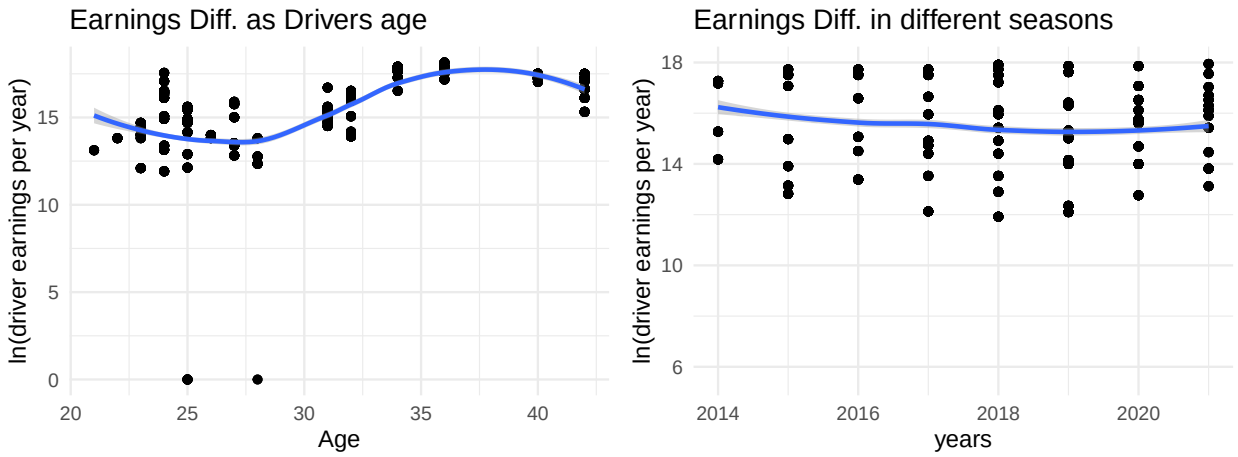
After the summary, I filtered for the drivers who raced in 2021 and for the hybrid era (2014-2021). The number of observation became 1547 and 23 variable. for all of our key variables At the merge of the F1 data frame and the new earnings data frame, I was careful to join them year by year, to not merge wrong values in the wrong place. Finally, some variables were created: Age variable from date of birth, log of the earnings per year to make analysis easier later.

Table 1: Models to uncover relation between log(wage) and other variables

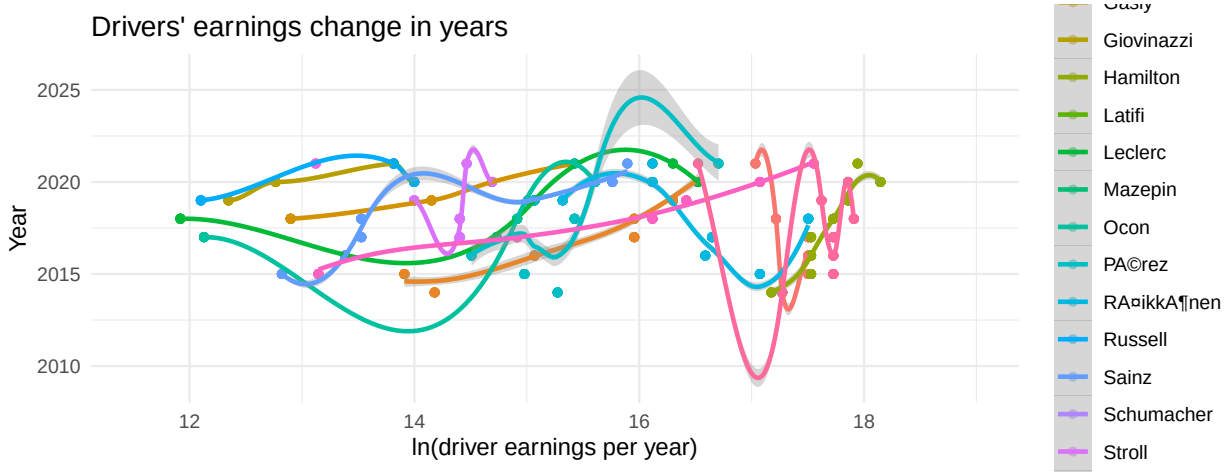
	(1)	(2)	(4)	(6)
(Intercept)	9.014***	9.207***	14.27***	9.413***
	(0.2544)	(0.2328)	(0.1622)	(0.2199)
age	0.2075***	0.1815***		0.1251***
	(0.0072)	(0.0062)		(0.0046)
p_earn_race_d		0.0868***		
		(0.0047)		
team_nameAlphaTauri			0.2061	1.683***
			(0.2173)	(0.1463)
team_nameAlpineF1Team			1.959***	2.333***
			(0.2503)	(0.0892)
team_nameAstonMartin			1.224***	2.133***
			(0.2929)	(0.1210)
team_nameFerrari			2.839***	3.111***
			(0.1685)	(0.1111)
team_nameForceIndia			0.2958	1.155***
			(0.1842)	(0.1069)
team_nameHaasF1Team			-0.4547**	1.030***
			(0.1622)	(0.1032)
team_nameManorMarussia			-14.27***	-13.18***
			(0.1622)	(0.0975)
team_nameMcLaren			2.475***	2.461***
			(0.1874)	(0.0845)
team_nameMercedes			2.809***	3.288***
			(0.1702)	(0.1220)
team_nameRacingPoint			0.5182**	1.659***
			(0.1762)	(0.1055)
team_nameRedBull			1.631***	3.091***
			(0.1970)	(0.1453)
team_nameRenault			0.0993	1.374***
.			(0.2266)	(0.2011)
team_nameSauber			-3.388***	-2.117**
.			(0.7223)	(0.7586)
team_nameToroRosso			-1.615***	-0.4053
.			(0.3043)	(0.2602)
team_nameWilliams			-0.2852	0.8013***
.			(0.1732)	(0.1056)
Team finish pos. after race				0.0628***
.				(0.0130)
Observations	1,547	1,547	1,547	1,547
R2	0.31035	0.39639	0.71038	0.76682

Analysis

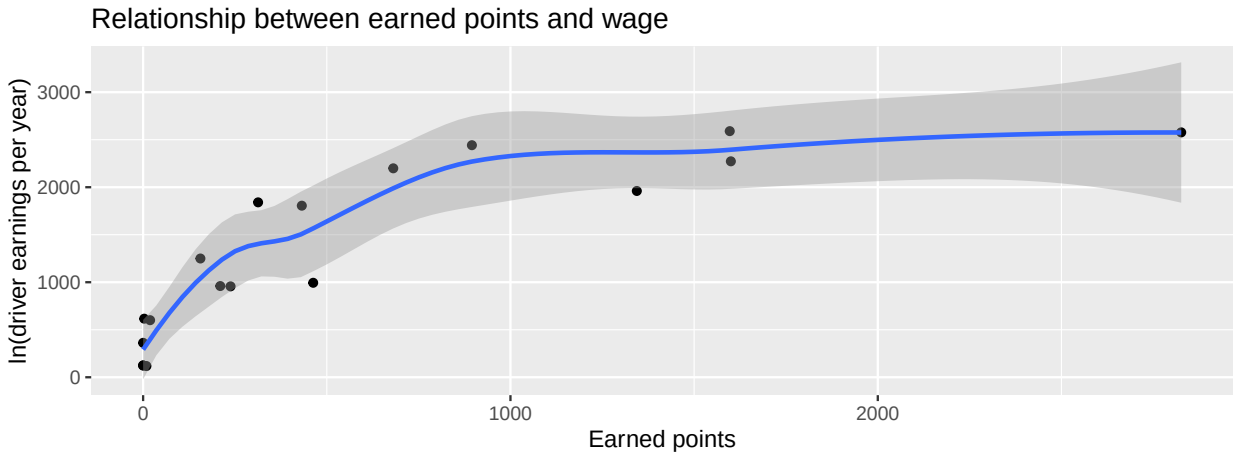
This assignment aimed to prove that the earned points or in other words the race results have a clear relationship with the reward the driver gets year by year. Interestingly the data shows different. The 3rd regression column's R2 presents that the relationship is too low. This is true for the 5th regression where the assumption is similar to the 3rd regression, but there is an extra control with the finish position of the team. This is a very interesting result, since in F1 the better a team finishes a season the more money they get back from the Formula 1 television rights. So, every team should focus on giving seat better drivers to reach an overall higher-finish score. Maybe a reason for that is the big difference in age between the drivers. The smaller teams are more likely to hire talented young drivers for less money. This strategy is risky since they need more attention and help, but in some cases, it has rewards like Max Verstappen and Lando Norris. Other teams hire the older drivers, who are more stable but demand much more salary. This is shown by regression 1. As the drivers' age increase by 1 year, they get 20.75% higher wages.



The log(wage) differences by age are also shown by this graph. After the young drivers join in F1 and spend a few years in the sport, they tend to earn less, since there is a lot of drivers who perform worse than the home team expects. The extreme values on the left side show the natural talents, who after a few years earn as much as world champion partners, like Verstappen. On the right side of the Graph, we can see a slight drop in the wage. A reason for that could be that the reflex and other skills become slower as they age and they perform worse. However, on the other second graph, we can see that the overall payment by the teams is quite constant, so the teams spend the same amount on drivers together. One reason for that they change the drivers between each other.



This graph shows us how the drivers' wages changed over the years. The dramatic changes are thanks to factors like retirement in the case of Alonso and team changes by the drivers. This is closely related to the 4th regression, which shows how the wages of the drivers change as they switch teams. The results of the regression are not so surprising. As the drivers switch to top teams like Mercedes AMG, Red Bull, McLaren, and Scuderia Ferrari, they earn sometimes nearly 3 times more than earlier. Small teams like Torro Rosso, Williams, Haas, and Marussia are the other way around, they pay 2-3 times less than the average.



Causal interpretation

As the last Graph show, everything has a top-end. The good talented drivers could reach their highest possible salary as they perform well at a small or middle team and tried to contract with one of the top constructors. Interestingly the position and the points the driver earns race by race is not that correlated with the wage of the drivers. Maybe this could be thanks to the fact that Mercedes Petronas AMG dominated the Hybrid era with 8 championship-winner cars with Lewis Hamilton. Or another reason could be the fact that Formula 1 is a technical sport. So, maybe the driver race constant on a high level, but the car is not capable of more. A good example is George Russel, who did an amazing job at Williams and from 2022 he will race with Hamilton at Mercedes.

Conclusion

After the analysis, we can say that in formula one money speaks. There are many factors we are not able to see, like how supported a driver is (how much extra money he can bring to the team), politics and others. A prove of that many talented children from smaller categories like Formula 2 and 3 could not race in higher categories, because they do not have the support, but they are on the podium all the time. This is sad, but not new in the motorsport world. It was and will be the entertainment of the wealthy people.

Appendix

Here you can see the additional charts

Chart1: Prove of data error

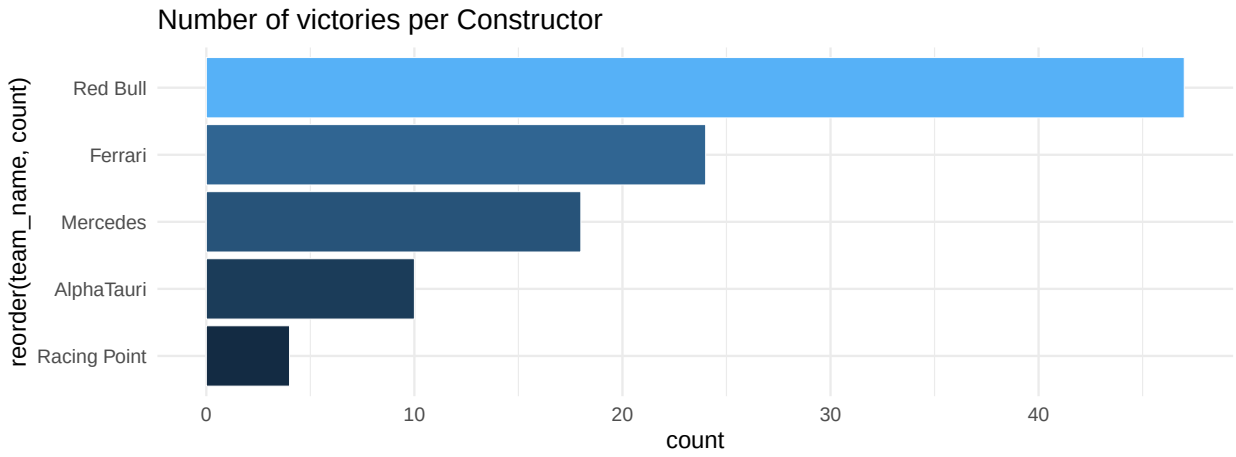


Chart2

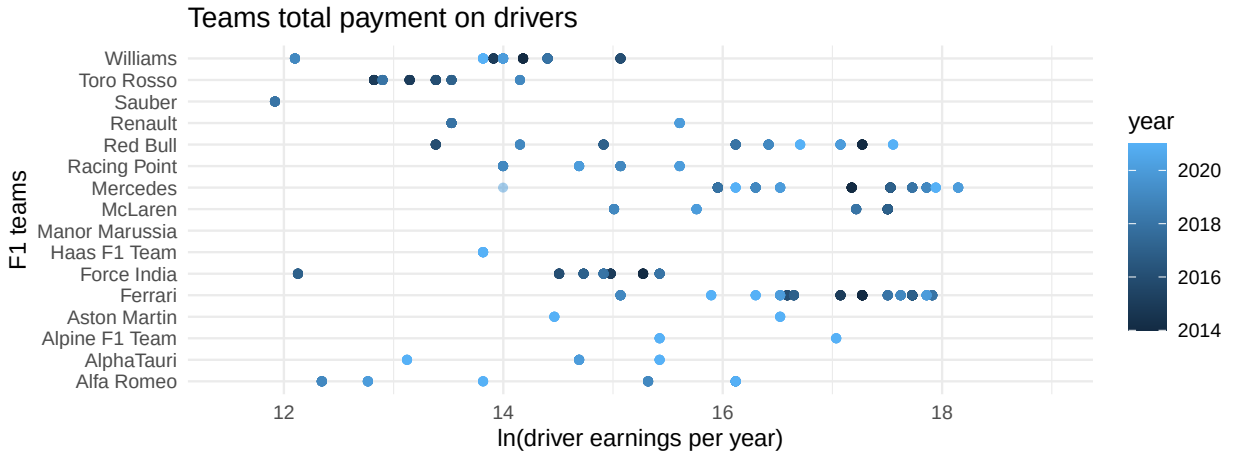


Chart3

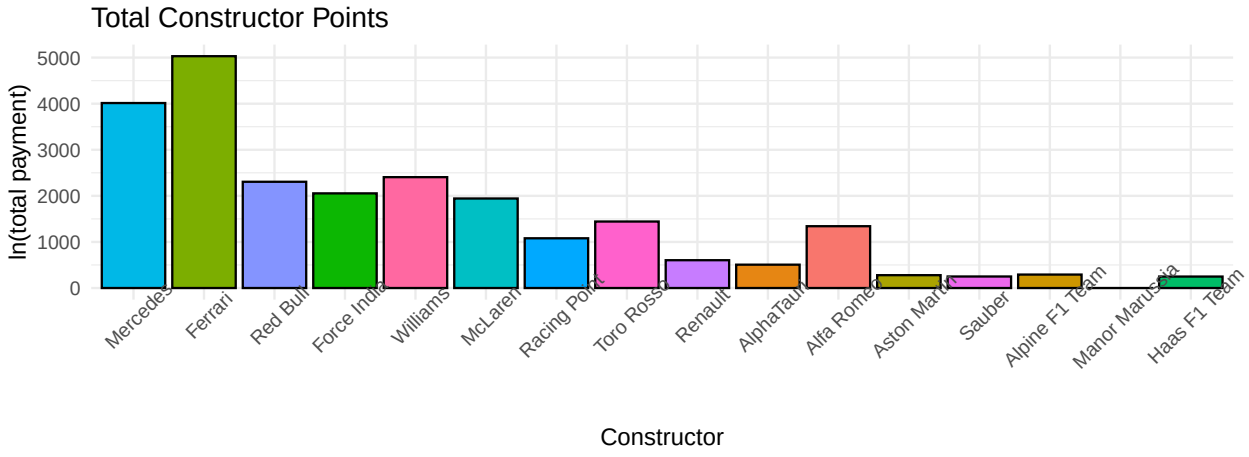


Chart4

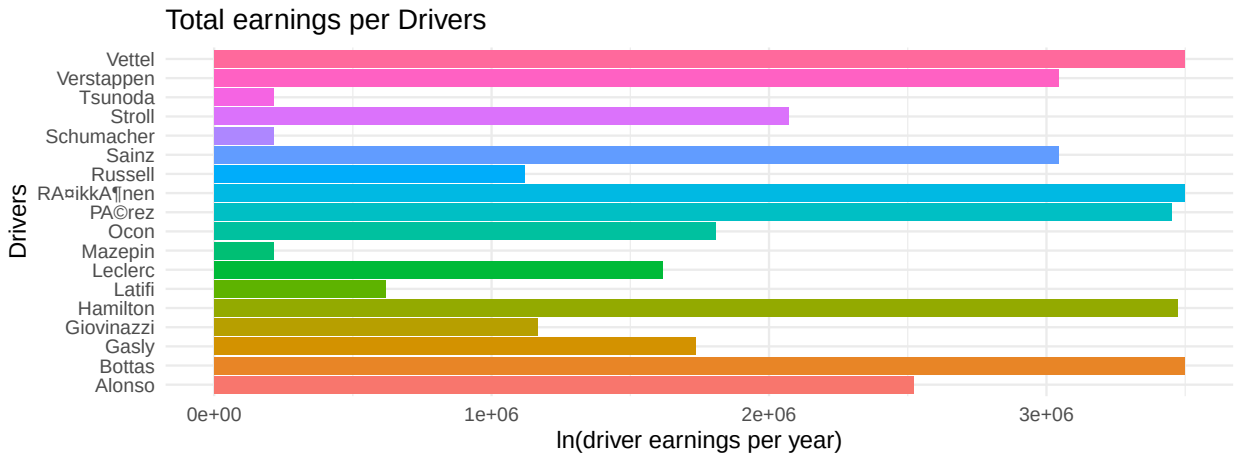


Chart5

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