LEARNING TO THINK OUT OF THE BOX

IBM Capstone Project

Intro

Do you ever think that someone could predict when and where there could be a car accident, and if the people involved would get hurt or won't? With data science, we can! And that's what we'll be doing in this project. It's just a little of common sense to think of the **reasons that could cause a car accident: the road conditions, the weather, the signaling, the stupidity of the driver**. But **do these factors really influence the chances of having a car accident? or are they just related with the severity of the accident itself?** For this project, we'll be using the **Seattle's report of collisions from 2004 till today** (you can click here to download it: Collision data from Seattle). We'll focus just in a some info of this dataset:

- **Collision address type**: was it on an intersection? on an alley? or just on a block?
- **Location**: it's very important for this project, to know if the accident was on a bridge, a tunnel, an avenue, etc.
- **Severity**: This data would be helpful to know the severity of the accident, if it resulted on just material damage or injuries.
- **Count of vehicles involved**: were there just two cars involved or was it a multiple colission?
- **Junction type**: With this, we'll obtain more info about the place where it happened.
- **Collision description and state code for that**: how was the accident? were two cars or was there a motocycle? did they collide on the left side or the right?
- **Weather**: Do the weather really influence the probabilities of having a traffic accident?
- **Road conditions**: How important it is if the road is wet or isn't, for example?
- **Light conditions**: Does it has anything to do if there it happens on a bright day or on a dark night? In the opposite of what we'll think at a first sight, We won't use the cases where there was speeding, because the obvious influence of this characteristic on the probabilities of having an accident.

We'll find the responses for this question on the next steps and we'll making an evaluation of it with **shiny colors and nice graphs** at the end. Thanks for reading!

Get on with it!

Ok. I've done it. I have two news. If you want to know the bad one first, read "Bad one" first, if you want to know the good one first, read "Good one" first.

Bad one: The results aren't what we expected.

Good one: Mmm... we... had... fun?

So, in my intro I told you that we would check if we could predict the severity of the accidents. Well, we can't. But I'll show you the analysis I did for finally understanding that the question we've must ask, was another one!

First of all, I dropped all the columns that weren't important for us, and just stay with the ones that would allow us to see the correlation between road conditions, severity, vehicles involved and weather. Then I started looking for relations between those characteristics:

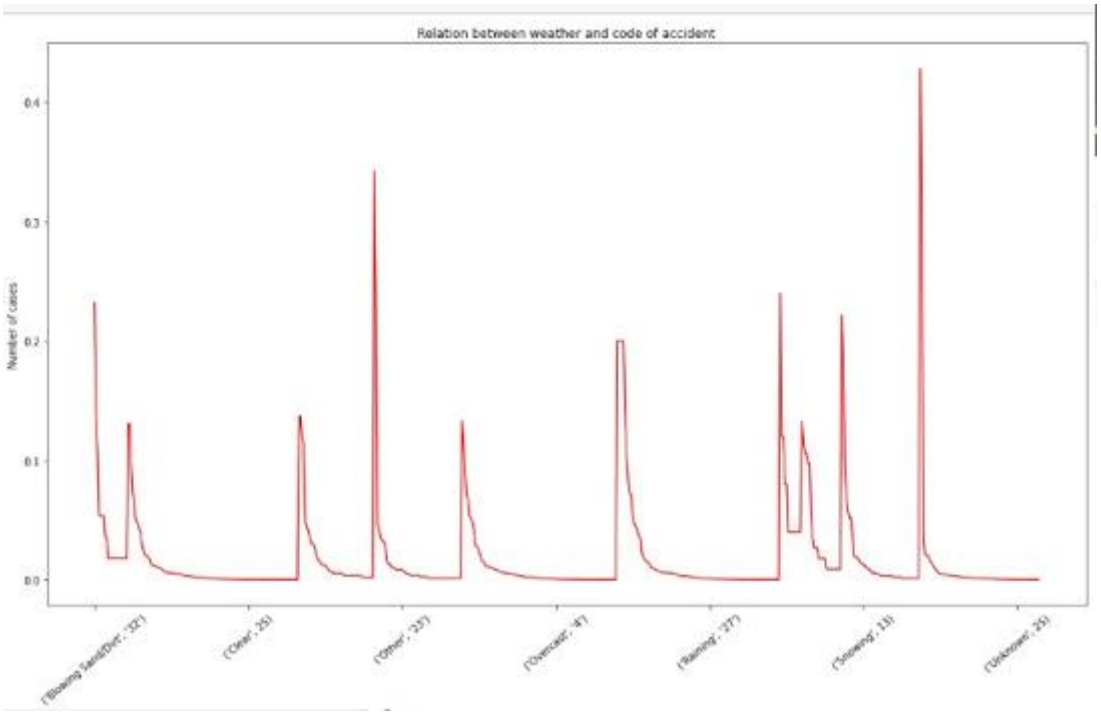
Clear	0.586180
Raining	0.174823
Overcast	0.146177
Unknown	0.079597
Snowing	0.004784
Other	0.004388
Fog/Smog/Smoke	0.003001
Sleet/Hail/Freezing Rain	0.000596
Blowing Sand/Dirt	0.000295
Severe Crosswind	0.000132
Partly Cloudy	0.000026

OK, most of days when there were accidents, were clear. That's unexpected.

But I'm pretty dyed-in-the-wool, so I'll insist on weather.

Let's remember our two severity code: code 1 is just for material damage, and code 2 is for injuries.

As we can check on the page of the references of the dataset we are using (click here), there are some codes that are used for knowing how was the accident. So, I took the most used of them and related them to the weather. And I get this:



What do we have here? Some issues: the most important, for me, is that Seattle is a very strange place where 'Blowing sand/Dirt' is more common that what we'd wish. Second, and a little less important, unless you do care about the data science aspect, is that those blowing

sand episodes, derivate on most collisions of "One Parked - One Moving" kind, and snowing ends up on collisions of two cars going in the same direction.

This isn't helpful, not a bit.

There has to be SOMETHING. I made myself a cup of tea and look at the microwave thinking of how a person gets hurt in a car collision. If someone sees me from the outside, I might be thinking of the existence of the world and solving my life for the next 40 years. But I'm not, I'm just looking at my tea spinning and thinking about cars and explosions.

I got an idea. If I'm driving and I, unfortunely, have an accident, I wouldn't like to be on a bridge or on a tunnel, because it sounds much more dangerous. It's obvious. It's going to be the answer. So I write the code and...

```
SEVERITYCODE  LOCATION
1             BATTERY ST TUNNEL NB BETWEEN ALASKAN WY VI NB AND AURORA AVE N          0.001477
              BATTERY ST TUNNEL SB BETWEEN AURORA AVE N AND ALASKAN WY VI SB          0.001300
              ALASKAN WY VI NB BETWEEN S ROYAL BROUGHAM WAY ON RP AND SENECA ST OFF RP 0.001305
              N NORTHGATE WAY BETWEEN MERIDIAN AVE N AND CORLISS AVE N                 0.001275
              AURORA AVE N BETWEEN N 130TH ST AND N 135TH ST                          0.001126
                                                                                        ...
2             YALE AVE N BETWEEN THOMAS ST AND HARRISON ST                            0.000017
              YESLER WAY BETWEEN JAMES ST AND OCCIDENTAL AVE S                        0.000017
              YESLER WAY BETWEEN WESTERN AVE AND POST AVE                             0.000017
              YORK RD S BETWEEN 36TH AVE S AND 37TH AVE S                             0.000017
              YORK RD S BETWEEN S HORTON ST AND 36TH AVE S                            0.000017
Name: LOCATION, Length: 35001, dtype: float64
```

So, if you are driving and you are so stupid to, I don't know, use your cellphone while you are driving, and that makes you crash with the car in front of you, and you are on an tunnel, take it easy because you won't hurt.

I didn't think I'd get mad for people not getting hurt on an accident, but I'm started to feel a little frustrated.

So, the severity isn't involved with the place neither. Amazing, I'm feeling my chakras destabilizing.

Maybe... it changes if it's on a intersection or on block?

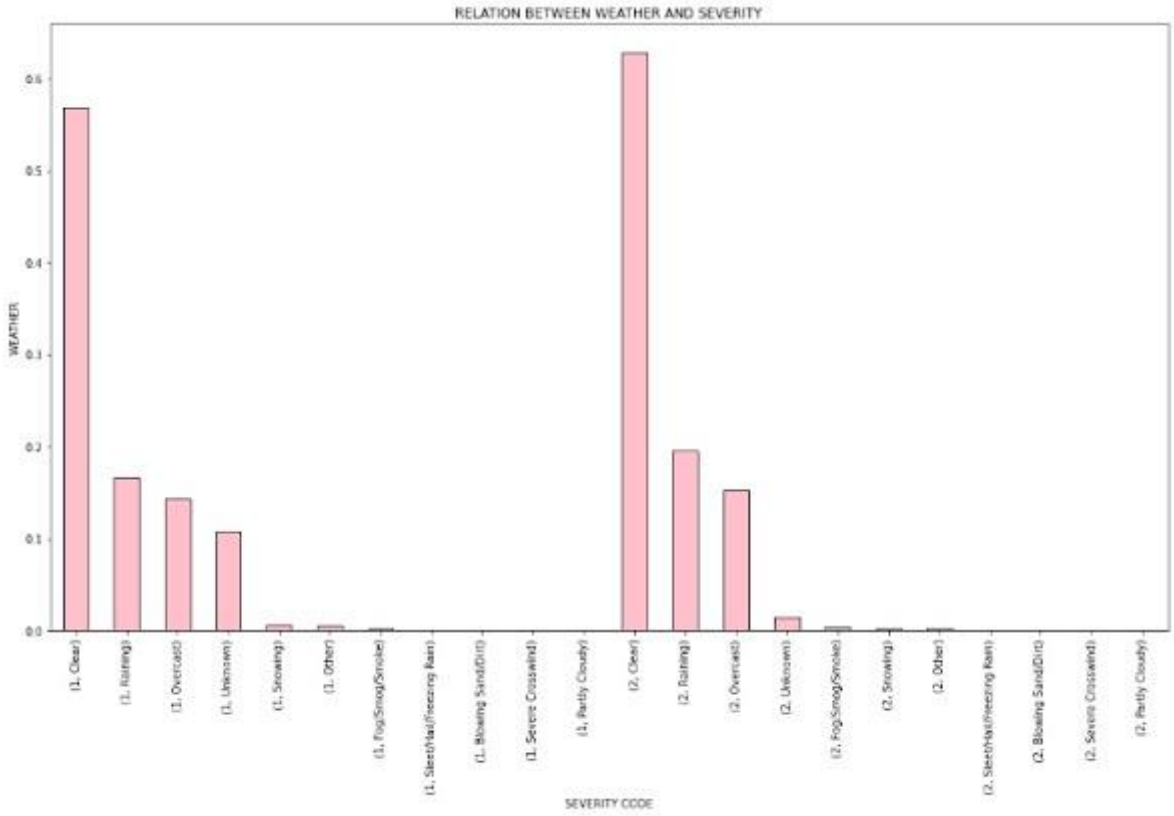
```
SEVERITYCODE  JUNCTIONTYPE
1             Mid-Block (not related to intersection)              0.539202
              At Intersection (intersection related)               0.272956
              Mid-Block (but intersection related)                 0.118669
              Driveway Junction                                    0.056964
              At Intersection (but not related to intersection)    0.011298
              Ramp Junction                                        0.000858
              Unknown                                              0.000054
2             At Intersection (intersection related)               0.470236
              Mid-Block (not related to intersection)              0.335779
              Mid-Block (but intersection related)                 0.126272
              Driveway Junction                                    0.055963
              At Intersection (but not related to intersection)    0.010781
              Ramp Junction                                        0.000934
              Unknown                                              0.000035
Name: JUNCTIONTYPE, dtype: float64
```

Ha! Okey. We got something here, but not something we couldn't have known with common sense. Most of accidents on mid-block were just material damage, and most of the crashes with injuries were at intersections. That's something.

With this little change on my investigation, I returned to the weather theme. I refuse to think that weather has nothing to do on it.

So I do all my best to make a beautiful bar plot and I choose my favourite color, pink, because I know this is going to be my lucky moment. Yes, I'm so inspired, I'm going to get it this time, and it's going to look georgous.

The result:



This was me before and after seeing the results of my bar plot:

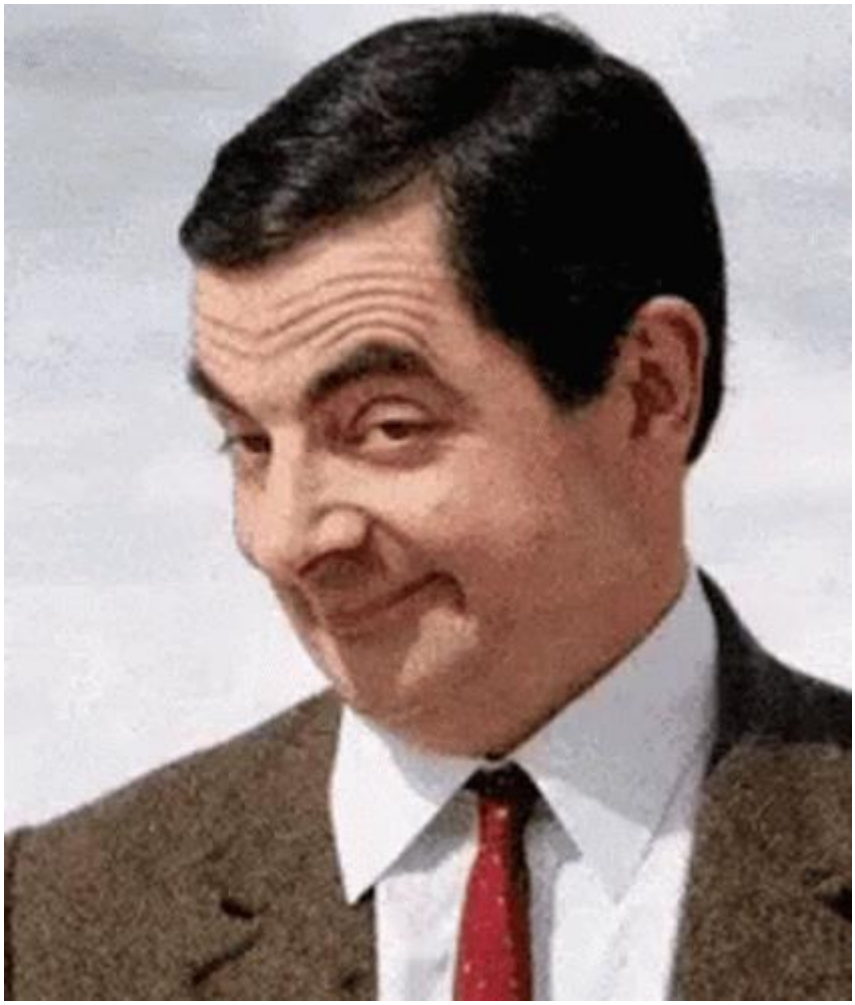
Most accidents with code 2 where on sunny days, it's obvious. I'm kidding, it's not obvious, it's the opposite of obvious, in fact.

And that was the moment, before throwing my laptop of the 5th floor, that I figured out I was asking the wrong question. What if we can't predict the factors that arrive to a more severe accident, because it deppends on human behaviour? what if the thing we could really predict, is where we are more exposed to be on a traffic accident?

Looking at my failed bar plot, I decided to look it in another way. So... I googled how many days does it rain annualy on Seattle: believe or not, it rains around 150 days per year.

Ok, in Seattle it rains a lot, but we have data from 2004 to today, so, asumming that there was 150 rainny days per year, counting them for the 16 years we have our dataset, there were 3440 sunny days and 2400 rainy days. A difference of 1000 days explains easily why were the biggest amount of accidents on sunny days.

And then... it happened THE MAGIC.



I figured out that there was something really obvious I wasn't seeing: the top ten of days with most accidents.

```
2006/11/02 00:00:00+00    96
2008/10/03 00:00:00+00    92
2005/05/18 00:00:00+00    84
2005/11/05 00:00:00+00    83
2006/01/13 00:00:00+00    83
2008/10/31 00:00:00+00    82
2005/04/29 00:00:00+00    76
2005/04/15 00:00:00+00    75
2004/12/04 00:00:00+00    74
2007/10/19 00:00:00+00    74
```

People from Seattle: HOW CAN YOU CRASH 96 TIMES ON A DAY?

Do you know what did these 10 days had in common?

IT WAS RAINING.

I could lie to you and say I figured it out doing great codes on Python, but the truth is I'm not very comfortable yet with coding and so I first went to a website that's called "Find out what the weather was like outside the day you were born!". Yeap, I was born in 10 different days, please don't make questions about it.

I entered the 10 days and I always got the same answer:

Once I got that, I checked on our database, but I have to be honest: I trust more in "weather sumofus" than in the Seattle's traffic department guys. I'm sorry if anyone of them is reading this. (I'm really not)

```
In [18]: eureka = df1.loc[df['INCDATE'] == '2006/11/02 00:00:00+00']
         eureka.head(50)
```

36956	1	Block	S SPOKANE NR ST BETWEEN 4TH AVE S AND 5TH AVE S	Property Damage Only Collision	2	2006/11/02 00:00:00+00	11/2/2006	Mid-Block (but intersection related)	MOTOR VEHICLE STRUCK MOTOR VEHICLE, REAR END	Raining
36995	1	Alley	NaN	Property Damage Only Collision	1	2006/11/02 00:00:00+00	11/2/2006	Mid-Block (not related to intersection)	NOT ENOUGH INFORMATION / NOT APPLICABLE	Unknown
37203	2	Block	DEXTER AVE N BETWEEN HALLADAY ST AND 4TH S AVE N	Injury Collision	2	2006/11/02 00:00:00+00	11/2/2006	Mid-Block (not related to intersection)	MOTOR VEHICLE STRUCK MOTOR VEHICLE, FRONT END ...	Raining
37231	1	Intersection	12TH AVE S AND S CHARLES S ST	Property Damage Only Collision	2	2006/11/02 00:00:00+00	11/2/2006	At Intersection (intersection related)	MOTOR VEHICLE STRUCK MOTOR VEHICLE, FRONT END ...	Raining

```
In [19]: eureka = df1.loc[df['INCDATE'] == '2005/05/18 00:00:00+00']
         eureka.head(50)
Out[19]:
```

	SEVERITYCODE	ADDRTYPE	LOCATION	SEVERITYDESC	VEHCOUNT	INCDATE	INCDTTM	JUNCTIONTYPE	SDOT_COLDESC	WEATHER	RI
313	1	Block	49TH AVE NE BETWEEN DEAD END 1 AND NE 39TH W ST	Property Damage Only Collision	2	2005/05/18 00:00:00+00	5/18/2005 2:00:00 AM	Mid-Block (not related to intersection)	MOTOR VEHICLE STRUCK MOTOR VEHICLE, LEFT SIDE ...	Raining	
1013	1	Block	3RD AVE BETWEEN VIRGINIA ST AND LENORA ST	Property Damage Only Collision	2	2005/05/18 00:00:00+00	5/18/2005 10:38:00 PM	Mid-Block (but intersection related)	MOTOR VEHICLE STRUCK MOTOR VEHICLE, REAR END	Overcast	

I told you we could trust my weird page!

Now I'm all motivated and want to check out the 10 days with less accidents.

Most of these days were on 2020, and in case you didn't hear: we are on the middle of a pandemic, so traffic on the last months wasn't like it is usually, and that made me look for days with less accidents outside of this disastrous year that I hope finishes very soon.

Guess what was the weather like on these days...

	SEVERITYCODE	ADDRTYPE	LOCATION	SEVERITYDESC	VEHCOUNT	INCDATE	INCDTTM	JUNCTIONTYPE	SDOT_COLDESC	WEATHER	ROA
192522	2	Intersection	ROOSEVELT WAY NE AND NE 43RD N ST	Injury Collision	1	2018/12/25 00:00:00+00	12/25/2018 6:53:00 PM	At Intersection (intersection related)	MOTOR VEHICLE STRUCK PEDESTRIAN	Overcast	
193289	1	Block	4TH AVE BETWEEN MARION ST AND MADISON ST	Property Damage Only Collision	2	2018/12/25 00:00:00+00	12/25/2018 6:38:00 AM	Mid-Block (not related to intersection)	MOTOR VEHICLE STRUCK MOTOR VEHICLE, FRONT END ...	Clear	
193451	2	Block	N 40TH ST BETWEEN ASHWORTH AVE N AND WOODLAWN ...	Injury Collision	2	2018/12/25 00:00:00+00	12/25/2018 2:37:00 AM	Mid-Block (not related to intersection)	MOTOR VEHICLE STRUCK MOTOR VEHICLE, REAR END	Clear	
193899	1	Intersection	5TH AVE AND SPRING ST	Property Damage Only Collision	2	2018/12/25 00:00:00+00	12/25/2018 7:32:00 PM	At Intersection (intersection related)	MOTOR VEHICLE STRUCK MOTOR VEHICLE, FRONT END ...	Clear	
194441	2	Block	12TH AVE BETWEEN E OLIVE ST AND E HOWELL ST	Injury Collision	3	2018/12/25 00:00:00+00	12/25/2018 10:07:00 PM	Driveway Junction	MOTOR VEHICLE STRUCK MOTOR VEHICLE, FRONT END ...	Clear	

Clear, of course. But I want you to pay attention specially on the months:

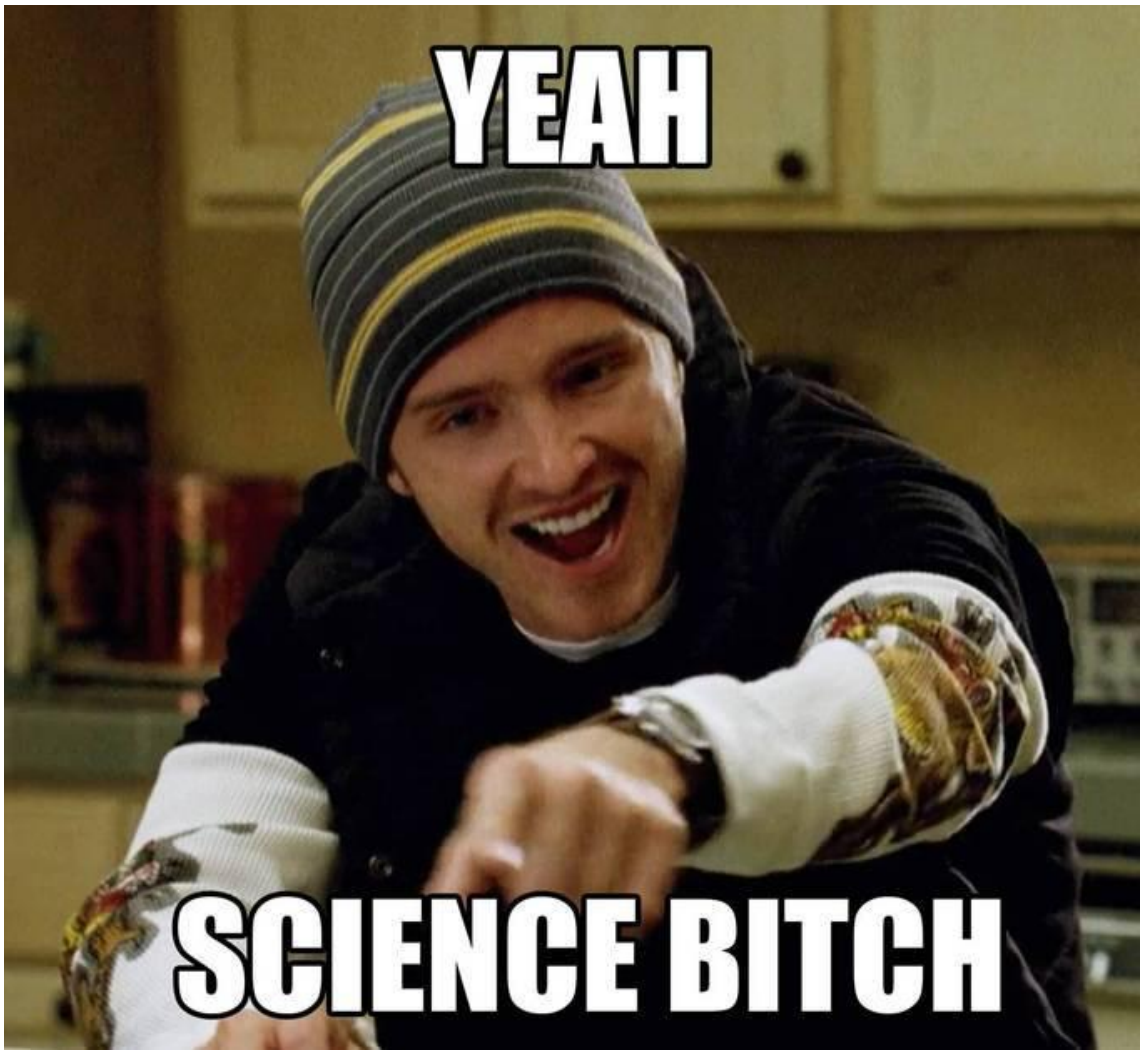
	Date	Amount of accidents
0	2010/11/25	8
1	2012/01/28	8
2	2005/04/26	8
3	2004/12/27	8
4	2009/12/06	8
5	2019/11/28	7
6	2010/12/26	7
7	2014/12/25	7
8	2004/12/12	7
9	2004/12/26	7

Instead of a clear day, do we see what do the most of these days have in common? they are most of them in November/December, and you know what happens on that season? IT'S F****** FREEZING, so people don't go out needlesly as they would on May or June.

My conclussion? First, as a data scientist, I learned my lesson: to correct the question instead of getting stuck with it even knowing it drives me at an endless point.

Second, severity has nothing to do with weather or the road conditions, but with the human action, so we cannot predict it, unless we do an IQ test to all the people that are trying to get a license, but that will be illegal and a little discriminatory.

Third, **most accidents happen on rainy days**. Yeap, this work was all to discover that. Welcome to my data science blog and drive carefuly if it's raining!



Note: This Project was written to be corrected by another student from the course, and it's been uploaded on my Blog, that's why it's so informal.