lightgray **Num**	**Family Type**	**Name**	**Label**	**Format**
1	ID	id_client	ID- Client ID	string
2	ID	id_vehicle	ID- Vehicle ID	string
3	ID	id_year	ID- Year	string
4	Claims	id_claim	Claims- Claim ID	string
5	Claims	claim_nb	Claims- Number of Claims	int
6	Claims	claim_amount	Claims- Total Claims Amount	int

Variables List : Claims database

lightgray **Num**	**Family Type**	**Name**	**Label**	**Format**
1	ID	id_client	ID- Client ID	string
2	ID	id_vehicle	ID- Vehicle ID	string
3	ID	id_policy	ID- Policy ID	string
4	ID	id_year	ID- Year	string
5	Policy	pol_bonus	Policy- Bonus Coefficient	float
6	Policy	pol_coverage	Policy- Coverage	string
7	Policy	pol_duration	Policy- Duration	int
8	Policy	pol_sit_duration	Policy- Current Endorsment Duration	int
9	Policy	pol_pay_freq	Policy- Payment Frequency	string
10	Policy	pol_payd	Policy- Payd Indicator	string
11	Policy	pol_usage	Policy- Usage	string
12	Policy	pol_insee_code	Policy- Insee Town Code	string
13	Drivers	drv_drv2	Drivers- Secondary Driver Presence Indicator	string
14	Drivers	drv_age1	Drivers- First Driver Age	int
15	Drivers	drv_age2	Drivers- Secondary Driver Age	int
16	Drivers	drv_sex1	Drivers- First Driver Gender	string
17	Drivers	drv_sex2	Drivers- Secondary Driver Gender	string
18	Drivers	drv_age_lic1	Drivers- First Driver Licence Age	int
19	Drivers	drv_age_lic2	Drivers- Secondary Driver Licence Age	int
20	Vehicle	vh_age	Vehicle- Vehicle Age	int
21	Vehicle	vh_cyl	Vehicle- Engine Capacity	int
22	Vehicle	vh_din	Vehicle- Din Power	int
23	Vehicle	vh_fuel	Vehicle- Fuel Type	string
24	Vehicle	vh_make	Vehicle- Make	string
25	Vehicle	vh_model	Vehicle- Model	string
26	Vehicle	vh_sale_begin	Vehicle- Sales Date Beginning	int
27	Vehicle	vh_sale_end	Vehicle- Sales Date End	int
28	Vehicle	vh_speed	Vehicle- Max Speed	int
29	Vehicle	vh_type	Vehicle- Type	string
30	Vehicle	vh_value	Vehicle- Value	int
31	Vehicle	vh_weight	Vehicle- Weight	int

Variables List : Underwriting database

4.1. **Variable id_client.** id_client is a string of the form Annnnnnnn ('A' followed by an 8-digit number). First client ID is A00000001 and last is A00091488. Why not A00100000 ? This is because a single client can own multiple vehicles, as we'll see in the next section.

4.2. **Variable id_vehicle.** id_vehicle as a string of the form Vnn (a 'V' followed by a 2-digit number). First vehicle is always numbered V01. If a client has multiple vehicles, then the numeration increases by 1. There is no particular ordering in the vehicles, so their rank should not represent anything valuable.

4.3. **Variable id_policy.** id_policy is a string of the form Annnnnnnn-Vnn, resulting from the concatenation of id_client, a minus sign, and id_vehicle. This is the unique ID that you must provide in you response CSV file, among with your calculated premium.

4.4. **Variable id_year.** Year ID begins at Year 0 and ends at Year 4. The Year ID is unique in each dataset.

Client ID, Vehicle ID and Year ID are present in the underwriting datasets (U_i) as well as in the claims datasets (C_i). When merging claims with contracts, don't forget to use the three IDs as keys.

4.5. **Variable pol_bonus.** The bonus/malus system is compulsary in France, but we will only use it here as a possible feature. The coefficient is attached to the driver. It starts at 1 for young drivers (i.e. first year of insurance). Then, every year without claim, the bonus decreases by 5% until it reaches its minimum of 0.5. Without any claim, the bonus evolution would then be : $1 \to 0.95 \to 0.9 \to 0.85 \to 0.8 \to 0.76 \to 0.72 \to 0.68 \to 0.64 \to 0.6 \to 0.57 \to 0.54 \to 0.51 \to 0.5$

Every time the driver causes a claim (only certain types of claims are taken into account), the coefficient increases by 25%, with a maximum of 3.5. Thus, the range of pol_bonus extends from 0.5 to 3.5 in the datasets.

4.6. **Variable pol_coverage.** The coverage are of 4 types : Mini, Median1, Median2 and Maxi, in this order. As you can guess, Mini policies covers only Third Party Liability claims, whereas Maxi policies covers all claims, including Damage, Theft, Windshield Breaking, Assistance, etc.

4.7. **Variable pol_duration.** Policy duration represents how old the policy is. It is expressed in year, accounted from the beginning of the current year i. Oldest policies in this portfolio can last since prehistoric ages of 45 years.

4.8. **Variable pol_sit_duration.** Situation duration represent how old the current policy caracteristics are. It can be different from pol_duration, because the same insurance policy could have evolved in the past (e.g. by changing coverage, or vehicle, or drivers, ...).

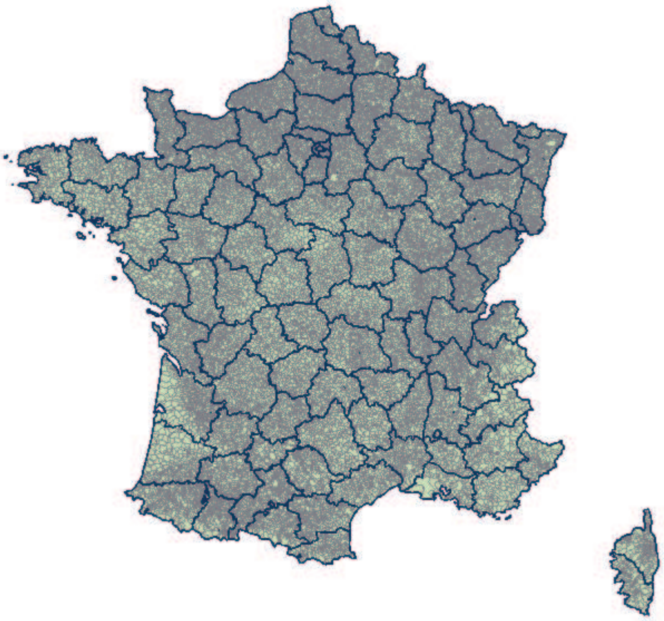
4.9. **Variable pol_pay_freq.** The price of the insurance coverage can be paid annually, bi-annually, quarterly or monthly. Be aware that you must provide a yearly cotation in your answer to the pricing game.

4.10. **Variable pol_payd.** The pol_payd is a boolean (i.e. a string with Yes or No), which indicates whether our client has subscribed a mileage-based policy or not. In those early ages of Year 0, Pay As You Drive was not that current, so they represent a minority in the portfolio.

4.11. **Variable pol_usage.** The policy use describes what usage the driver makes from his vehicle, most of time. There are 4 possible values : WorkPrivate which is the most common, Retired which is presumed to be aimed at retired people (who also are presumed driving less kilometers), Professional which denotes a professional usage of the vehicle, and AllTrips which is quite similar to Professional (including pro tours). As for the coverage, it would be very surprising that this variable had no effect on frequency...

4.12. **Variable pol_insee_code.** insee_code is a 5-digits alphanumeric code used by the French National Institute for Statistics and Economic Studies (hence INSEE) to identify communes and departments in France. There are about 36,000 'communes' in France, but not every one of them is present in the dataset (there are only 18,000 of them). The first 2 digits of insee_code identifies the 'department' (they are 96, not including overseas departments). The insee_code or department code can be used to possibly merge external data to the datasets : population density, OSM data, etc.

In case you need it, two shapefiles are available online : one DEPARTMENTS.zip for departments, and one COMMUNES.zip for communes. Be aware that, if you need to graph geographical information, french reference system is RGF93 / Lambert-93 (EPSG :2154) and not the common WGS84.
http://freakonometrics.free.fr/PG3/COMMUNES.zip
http://freakonometrics.free.fr/PG3/DEPARTEMENTS.zip



4.13. **Variable drv_drv2.** The drv_drv2 boolean (Yes/No) identifies the presence of a secondary driver on the vehicle. There is always a first driver, which characteristics (age, sex, licence) are provided, but a secondary driver is optional, and is present 1 time out of 3.

4.14. **Variable drv_age1.** This is quite obviously the age of the first driver. drv_age is expressed in years counted from the beginning of the considered year. Then, drv_age increases by 1 every year, like in real world... Legal age to drive is 18, so you shouldn't find any age below that limit. Due to the fact that the database is built on existing situations before Year 0, in fact the minimum age is 19 in Year 0 dataset. On the other side, you'll also find quite old drivers.

4.15. **Variable drv_age2.** When drv_drv2 is Yes, then the secondary driver's age is present. When not, this age is 0.

4.16. **Variable drv_sex1.** European rules force insurers to charge the same price for women and men. But driver's gender can still be used in academic studies, and that's why drv_sex1 is still available in the datasets, and can be used as discriminatory variable in this pricing game.

4.17. **Variable drv_sex2.** As for drv_sex1, drv_sex2 represents the gender of the optional secondary driver. You'll notice that the distribution of this variable is opposite to drv_sex1.

4.18. **Variable drv_age_lic1.** drv_age_lic1 is the age of the first driver's driving licence. As for the other ages, it is expressed in integer years from the beginning of the current year.

4.19. **Variable drv_age_lic2.** drv_age_lic2 is the age of the second driver's driving licence. Be cautious that there are some outliers in the dataset.

4.20. **Variable vh_age.** This variable is the vehicle's age, the difference between the year of release and the current year. One can consider that vh_age of 1 or 2 correspond to new vehicles.

4.21. **Variable vh_cyl.** The engine cylinder displacement is expressed in ml in a continuous scale. This variable should be highly correlated with din power of the vehicle.

4.22. **Variable vh_din.** The vh_din is a representation of the motor power. Don't be surprised to find correlations between din power, cylinder, speed and even value of the vehicle...

4.23. **Variable vh_fuel.** vh_fuel is the motor alimentation, with mainly two values Diesel and Gasoline. Very few Hybrid vehicles can also be found, but, 6 years ago, the hybrid market was still at its beginning.

4.24. **Variable vh_make.** The make (brand) of the vehicle. As the database is built from a french insurance, the three major brands are Renault, Peugeot and Citroën.

4.25. **Variable vh_model.** As a subdivision of the make, vehicle is identified by its model name. The are about 100 different make names in the datasets, and about 1,000 different models. Should you use them, consider concatenating vh_make and vh_model.

4.26. **Variables vh_sale_begin and vh_sale_end.** vh_sale_begin and vh_sale_end are the dates (in fact : ages) from the beginning of the current year of the beginning and the end of marketing years of the vehicle. This could for instance identify policies that covers very new vehicles or second-hand ones.

4.27. **Variable vh_speed.** This is the maximum speed of the vehicle, as stated by the manufacturer.

4.28. **Variable vh_type.** vh_type can be Tourism or Commercial. You'll find more Commercial types for Professional policy usage than for WorkPrivate.

4.29. **Variable vh_value.** The vehicle's value (replacement value) is expressed in euros, without inflation so it should be stable from a year to another.

4.30. **Variable vh_weight.** vh_weight is the weight (in kg) of the vehicle.

4.31. **Variable id_claim.** As the claims datasets PG_2017_CLAIMS_YEARi.CSV shows individual claims, we should be able to identify them. id_claim is a string of the form CLnn (CL followed by a 2-digit number). Numbering of the claims begins at 1 for every policy and each year. Then, the last value of id_claim is the maximum number of claims for a vehicle in a year. Two-digits representation is sufficient : this maximum doesn't exceed 7 (but not on Year 0, where the maximum is 6).

4.32. **Variable claim_nb.** As we are talking about individual claims, each claim_nb has a value of 1. This variable is present for commodity purpose : this is the one you'll probably want to model in a frequency approach.

4.33. **Variable claim_amount.** Individual claim amounts range from (approx.) -2,000 to +300,000. Yes, there are negative values, they come from claims where our driver's liability is not engaged, so there's a legal recourse.