





Data dictionary

 Column	 Description
<u>id_policy</u>	Unique policy identifier
<u>year</u>	This dataset covers a 5 year range. This indicates the year of the contract <code>1-5</code> .
<u>pol_no_claims_discount</u>	This is a compulsory system in most European countries and globally. The coefficient is attached to the driver and indicates previous claims. It starts at <code>0.631</code> for new drivers (i.e. first year of insurance). Then, every year without claim, the value decreases by <code>0.047</code> until it reaches its minimum of <code>0</code> . Without any claim, the value evolution would then be: <code>0.631</code> , <code>0.584</code> , <code>0.538</code> , <code>0.491</code> , ... , <code>0.024</code> , <code>0</code> . Every time the driver causes a claim (only certain types of claims are taken into account), the coefficient increases by <code>0.203</code> , with a maximum of <code>1</code> . Thus, the range of <code>pol_no_claims_discount</code> extends from <code>0</code> to <code>1</code> in the dataset.
<u>pol_coverage</u>	There are four types of coverage: <code>Min</code> , <code>Med1</code> , <code>Med2</code> and, <code>Max</code> , in this order. <code>Min</code> policies cover only <u>Third Party Liability</u> claims, whereas <code>Max</code> policies covers all claims, including Damage, Theft, Windshield Breaking, Assistance, etc. The two <code>Med</code> policies represent intermediate coverage.
<u>pol_duration</u>	Policy duration represents how old the policy is. It is expressed in years, accounted from the beginning of the current year. There may be very old policies in the dataset for loyal customers.
<u>pol_sit_duration</u>	Represents how old the current policy characteristics are. This is different from <code>pol_duration</code> , because the same insurance policy could have evolved in the past (e.g. coverage, or vehicle, or drivers, ...). The <code>pol_sit_duration</code> is a metric for how up-to-date the characteristics are.
<u>pol_pay_frequency</u>	The price of the insurance coverage can be paid annually, bi-annually, quarterly or monthly. Be aware that you must provide a yearly quotation in your final submission.
<u>pol_payd</u>	PAYD or "Pay As You Drive" is a boolean (i.e. a string with <code>Yes</code> or <code>No</code>), which indicates whether the client has subscribed a mileage-based policy or not.
<u>pol_usage</u>	This indicates how the driver uses the vehicle with 4 possible values: <code>WorkPrivate</code> , <code>Retired</code> , <code>Professional</code> (which denotes a professional usage of the vehicle), and <code>AllTrips</code> which is quite similar to <code>Professional</code> (including professional tours).
<u>drv_sex1</u>	European rules force insurers to charge the same price for women and men. But gender can still be used in academic studies, and that's why <code>drv_sex1</code> is still available in the datasets. For our purposes, let's assume that this can be used as variable in this pricing game. In this dataset the values are <code>M</code> or <code>F</code> .
<u>drv_age1</u>	This is the age of the first driver in years.
<u>drv_age_lic1</u>	The age of the first driver's driving license in years.
<u>drv_drv2</u>	This boolean (<code>Yes</code> / <code>No</code>) identifies the presence of a secondary driver on the contract. There is always a first driver, for whom certain characteristics are provided, but a secondary driver is optional.
<u>drv_sex2</u>	The gender of the second driver if present in the data.
<u>drv_age2</u>	When <code>drv_drv2</code> is Yes, then the secondary driver's age is present. When no second driver is registered, this is <code>0</code> .
<u>drv_age_lic2</u>	The age of the second driver's driving license in years, if present.
<u>vh_make_model</u>	A hash representing the make and model of the vehicle. These hashes (random strings) represent text such as <code>Ford Focus</code> or <code>Honda Civic</code> .
<u>vh_age</u>	This variable is the vehicle's age, the difference between the year of production of the vehicle and the current year. One can consider values of <code>1</code> or <code>2</code> to correspond to new vehicles.
<u>vh_fuel</u>	<code>Diesel</code> , <code>Gasoline</code> or <code>Hybrid</code> . This is the fuel type of the vehicle.
<u>vh_type</u>	<code>Tourism</code> or <code>Commercial</code> . You'll find more <code>Commercial</code> types for <code>Professional</code> <code>policy_usage</code> than for <code>WorkPrivate</code> .
<u>vh_speed</u>	This is the maximum speed of the vehicle, as stated by the manufacturer.
<u>vh_value</u>	The vehicle's value (replacement value) is expressed in Euros.
<u>vh_weight</u>	The weight (in kg) of the vehicle.
<u>population</u>	The population of the town where the policy holder is registered multiplied by a constant scale.
<u>town_surface_area</u>	Approximate surface area (in hectors multiplied by a constant scale.) of the town where the policy holder is registered.
<u>claim_amount</u>	The amount of a claim expressed in euros for the year (a value of zero indicates that no claim was made).