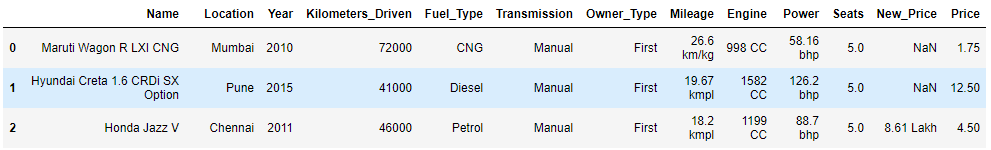
# Prieskumná analýza dát

Na projekt sme vybrali dáta o ojazdených autách z webu kaggle.com. Link: <https://www.kaggle.com/avikasliwal/used-cars-price-prediction> Dáta obsahujú informácie o názve auta, mieste, kde sa predáva auto, rok výroby, najazdené kilometre, typ pohonnej hmoty pre auto, typ prevodovky, koľko majiteľov malo auto, spotreba auta, zdvihový objem motora, výkon vozidla, počet sedadiel, cena nového vozidla, cena vozidla.

Ukážka prvých troch riadkov dát:

## Analýza dát

Dáta obsahujú 6019 riadkov záznamov. V 5-tich stĺpcoch z 13 máme nulové hodnoty. Na základe ďalšej analýzy sa rozhodneme, či a ako budeme dané hodnoty dopĺňať. Ak sa pozrieme na unikátne hodnoty v stĺpcoch, tak ich počty sú celkovo nevýrazné, u väčšiny v rozmedzí <2,540>. Tri stĺpce majú výrazne viac unikátnych hodnôt, čo ale vychádza z podstaty ich informácií (názov auta, najazdené kilometre, cena).

Skontrolovaním počtu hodnôt v jednotlivých stĺpcoch sme identifikovali 2 problémy. Počet sedadiel je uvedený v desatiných číslach a jedno vozidlo ma udaný počet miest na sedenie 0. Toto auto sme v dátach vyhľadali a podľa zadaného mena ručne upravíme počet sedadiel.

Stĺpce s názvom vozidla a miestom predaja dáme na malé písmená. Pre prípad chyby pri vytváraní tabuľky.

Stĺpec s údajmi o roku výroby vozidla je bez chýbajúcich hodnôt s pár outliermi na základe box plotu. Týchto pár hodnôt, ale nepovažujeme za outliery, nakoľko sa pár starších vozidiel vyskytovať môže.

Číselný stĺpec s údajmi o najazdených kilometroch už nejaké tie vychýlene hodnoty obsahuje. Nespochybniteľne nesprávna hodnota je uvedená pri jednom vozidle, kde vozidlo vyrobené v roku 2017 ma najazdených 6 500 000km. **Ako toto fixneme? asi deleat ci?** V stĺpci sa nachádzajú aj ďalšie vychýlené hodnoty, ale podľa ručného skontrolovania, tu už je pravdepodobnejšie, že tieto autá mohli tieto kilometre najazdiť napr. druhá najväčšia najazdená vzdialenosť je 775000 a auto bolo vyrobené v roku 2013 ==**neviem ci to chceme takto zdovodnovat alebo mazeme outliers :)**

V stĺpcoch o type paliva, prevodovky a majiteľovi nevidíme žiadnu chybu.