

# Exotic Car Hack

- A. Sai Gowtham Babu

## Objective: -

Car prices have always been fluctuating and most people correctly don't know how to sell or buy a new car. But there are some car dealers / auto investors who used this term called as "CAR HACK" to buy/ sell a car. The meaning of this term is – you buy a car for a low price and sell in for a high price (gaining the profit from the investment). This is also be helpful when you want to buy a car for yourself. The main goal of this project is to *design a system that can car hack every car present in the market and let you know weather buying that car is a legal steal/ make you bankrupt*. This technique is more often used for buying exotic cars (sports cars, luxury sedans, SUV's, MUV's) since they are priced at high price and people often can't correctly judge the price of them. In this project, I am hacking my favourite car Ford Mustang GT (all models, years and options).

There are two options available in this: -

- 1) For buying: - A user can input the details of car he/she want to buy and their current pin code/ location. This allows the user to choose the number of miles he wants the car to be so that he can go and pick it up.
- 2) For selling: - A user can get an estimate price of their car and sell it based on the estimate (price is quoted based on the year, model, mpg, colour).

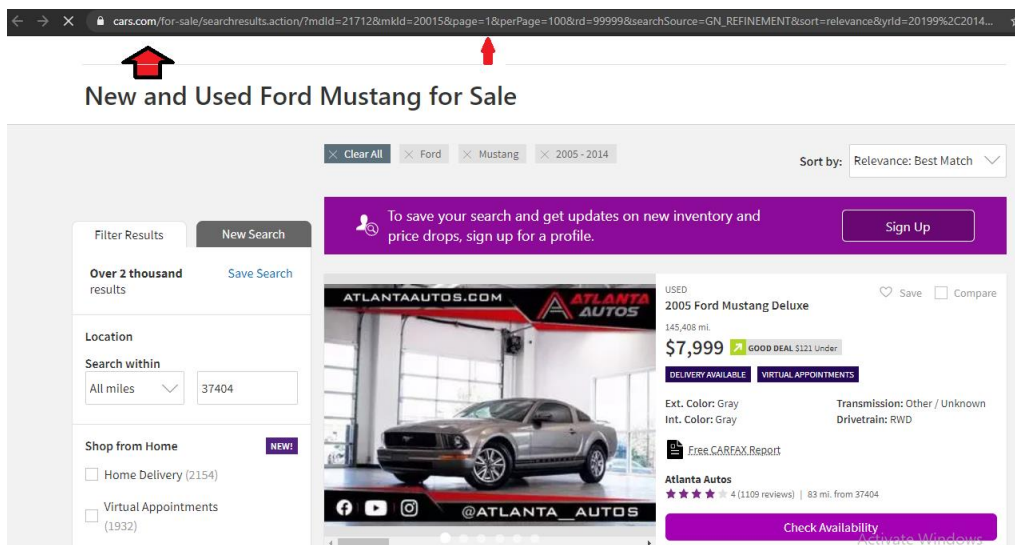
## Source credits: -

All the credits go to cars.com for having such a wonderful debase of each and every car. This helped me to collect all the data that I needed. Thank you @cars.com

## Data Collection: -

### Website link: -

For the data collection process, we need to find the location where the data is present and the unique value all the cars have. The picture below will make it clear for you.

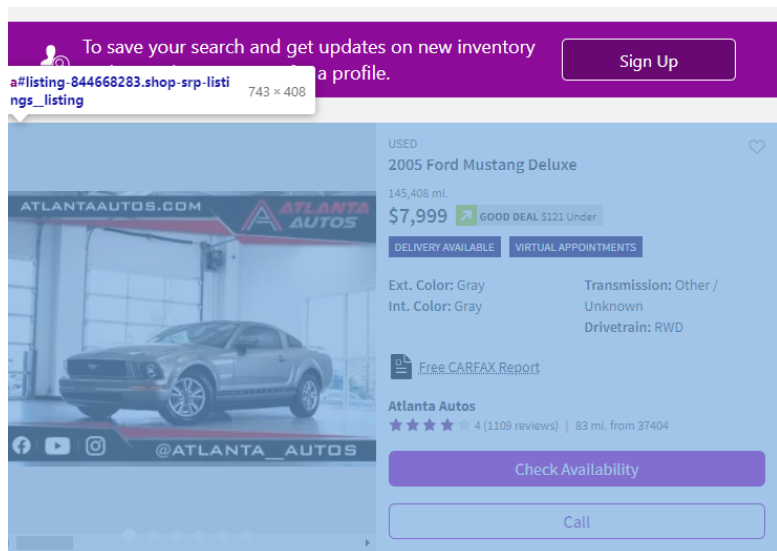


The link is shown in the big red arrow and this page contains 100 mustang cars in the year between 2005 and 2015. If you can see, there is another small arrow in the page. This lets us iterate through all the pages that is present in the given model years.

### Unique Values: -

Now, the question is how can I get the unique vehicle id for each vehicle?

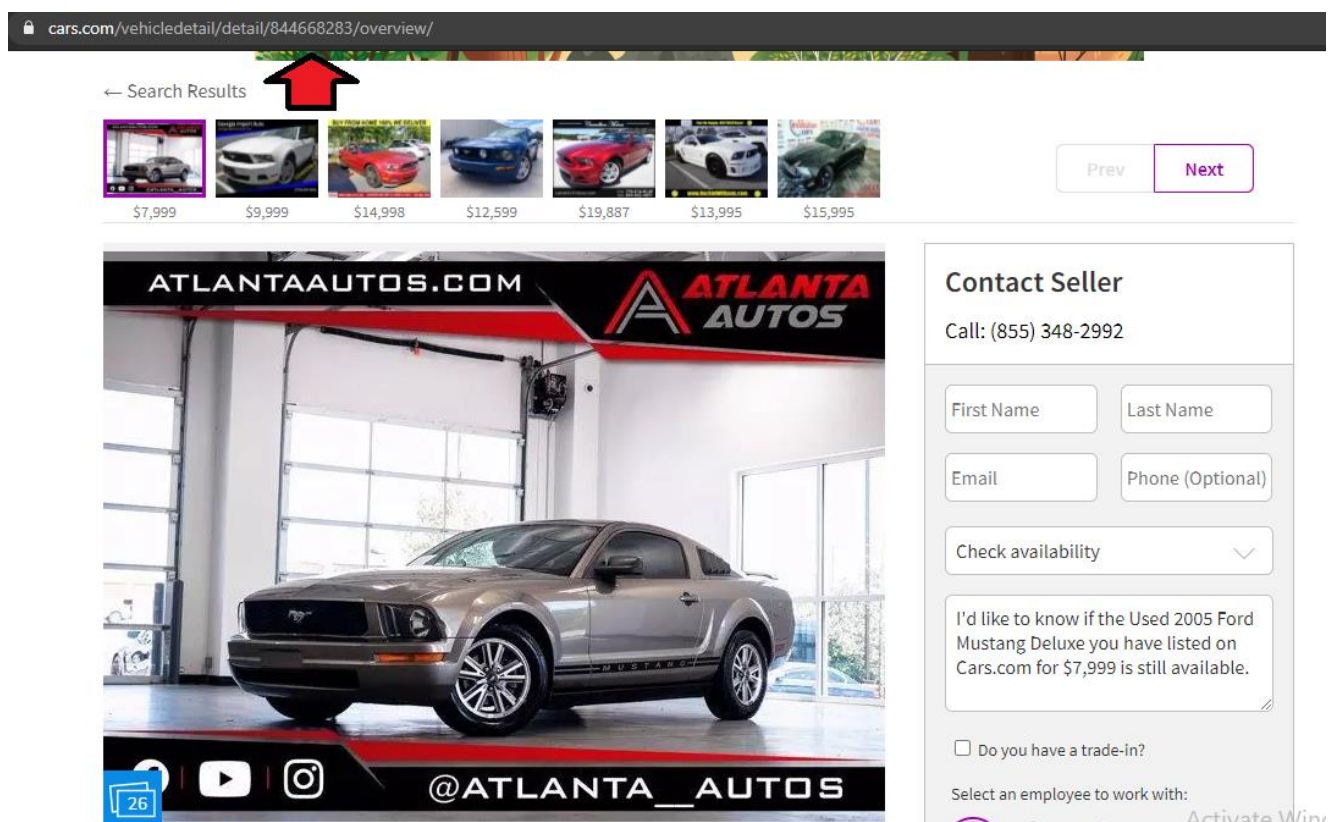
It is simple to scrape that in the main webpage itself. When we inspect the single car present in the page, we can see a href link that contains the id and the link to particular car. We can scrape that link for all the 100 cars present in a single page. This makes our process simple.



```
<script type="application/json" id="vehicleItemListSchema">...</script>
<script type="application/json">...</script>
<div class="shop-srp-listings_listing-container">
  <a href="/vehicledetail/detail/844668283/overview/" target="_self"
    id="listing-844668283" class="shop-srp-listings_listing" data-ixt
    data-goto-vdp="844668283" data-position="1" data-linkname="md-thumb"
    aria-label="2005 Ford Mustang Deluxe"> == $0
  <div class="shop-srp-listings_inner">
    <div class="listing-row_photo-container">
      <div class="badge-wrapper">
        <div class="listing-row_photo shadowed">
          <div class="hover-overlay">
            <span class="arrow left" style="display:none;" data-click=
              "scroll-left"></span>
            <span class="arrow right" data-click="scroll-right"></span>
          </div>
          <div class="photo-scroll-wrapper"></div>
          <div class="nav-dots"></div>
        </div>
      </div>
    <div class="listing-row_details">
      <div class="listing-row_stocktype">
        <div>
          <div>
            Used
          </div>
          <h2 class="listing-row_title">
            2005 Ford Mustang Deluxe
          </h2>
        </div>
      </div>
    </div>
  </div>
</div>
```

### Getting Cars data: -

As seen above, after getting the unique values is quite simple from next on. We can iterate through all the links that we got, collect the data from the links and make it a data frame and then combine all the data frames to form a single file. The only task here is to collect the base link of each model.



## Result: -

The scraper has collected over 14,000 data points.

## Data Cleaning: -

### Unique Values: -

After the collection of data, it turns to be that there are only 9,000 unique values present in the data. Since, we have looped over many pages and written separate code for years category, the same car might have been repeated. It might be a common thing since we are getting data from a same car model. Further cleaning can also decrease the data points based on the quality of our data.

### Problem in different pages: -

Since, this is a website owned by cars.com, I thought all the pages are going to be same format. But, here almost 20% of the cars listed in the website are structured in different format. So, the machine almost scraped 2000 data points of dirty data (the features of the column classified/ labelled incorrectly). Deleting all the data is bad for our analysis. So, we try to do some manual work using excel and try to clean most of the data manually.

### Car\_name cleaning: -

Most part of the names are correctly classified. But, some of them are having a tag called as “certified” before them. So, for that reason we replace that certified with a space. (It’s not done by ctrl+f and replace. Separate technique is performed for this as the above command doesn’t support it.) Also, dividing the full name into separate components like Year, model and type.

Before: -

1	car_name
10	Certified 2013 Ford Mustang V6 Premium
12	Certified 2013 Ford Mustang V6 Premium
17	Certified 2014 Ford Mustang V6
544	Certified 2013 Ford Mustang V6 PREMIUM
590	Certified 2012 Ford Mustang V6
360	Certified 2013 Ford Mustang V6
363	Certified 2018 Ford Mustang EcoBoost Premium
387	Certified 2017 Ford Mustang V6
305	Certified 2018 Ford Mustang GT
306	Certified 2018 Ford Mustang ECOBOOST
315	Certified 2018 Ford Mustang GT Premium
330	Certified 2017 Ford Mustang EcoBoost

After: -

Year	Model	Type
2013	Mustang	Ecoboost
2014	Mustang	Ecoboost
2005	Mustang	Premium
2009	Mustang	GT
2014	Mustang	GT
2011	Mustang	GT
2013	Mustang	Ecoboost
2013	Mustang	Ecoboost
2013	Mustang	GT

### Type and model Cleaning: -

This field contains the Model name (mustang, Shelby) and Type (Eco boost, GT, premium, Bullit, GT500, GT350, boss, Mach-e, standard) after cleaning them. Before cleaning, this column contains names like 302, 429, 750, California, bullit, GT deluxe, premium, route, rousch, boss, anniversary, v6, cayote, eco. But based on the domain knowledge, I have classified them into main categories to eliminate more model types. For the models which I cannot combine, I removed them since there are not more than 5-10 cars present for a particular model (ex: - cayote swap, supercharged). This turned out, I lost over 30 data points.

### Price and mileage cleaning: -

Around 1000 data points are misclassified based on the data we have; (like the example shown below) we can fill in some of the variables using central tendency methods (its better to keep the data instead of removing all the points). I tried using joins and merging them in excel by creating different excel sheets. Also used VLOOKUP on multiple columns to extract the data. With all the techniques, I was only able to fill in about 640 data points. The remaining data points have to be removed since they are not possible to fill the values by any techniques.

car_nar	price	mileage	colour	city_mile	highway_mile
2006 Ford	-1	Not	17	25	25
2014 Ford	-1	Not	19	29	29
2005 Ford	-1	Not	19	28	28

### Colour:

Most of the car colours are misclassified/ not correctly identity by the machine (because of the problem I discussed above). So, I have filled the misclassified data with black colour since most of the cars in the website seems to be black in colour.

### Remaining columns: -

There is not cleaning required in the case of remaining columns, most of them good and just required white space trimming, column transformation and text replacement.

### Result: -

After the total data cleaning process, the end result was about 8,400 data points after cleaning of data. This might be bit bad loosing over 600 data points just in cleaning. But I have no other choice to remove the data instead of having bad quality data for my project. I think I can do better with these 8,400 points. Let's hope for the end results.