KIRILL KANSHIN, PHD

CASH4CLUNKERS

CRAIGSLIST POSTINGS QUALITY CONTROL

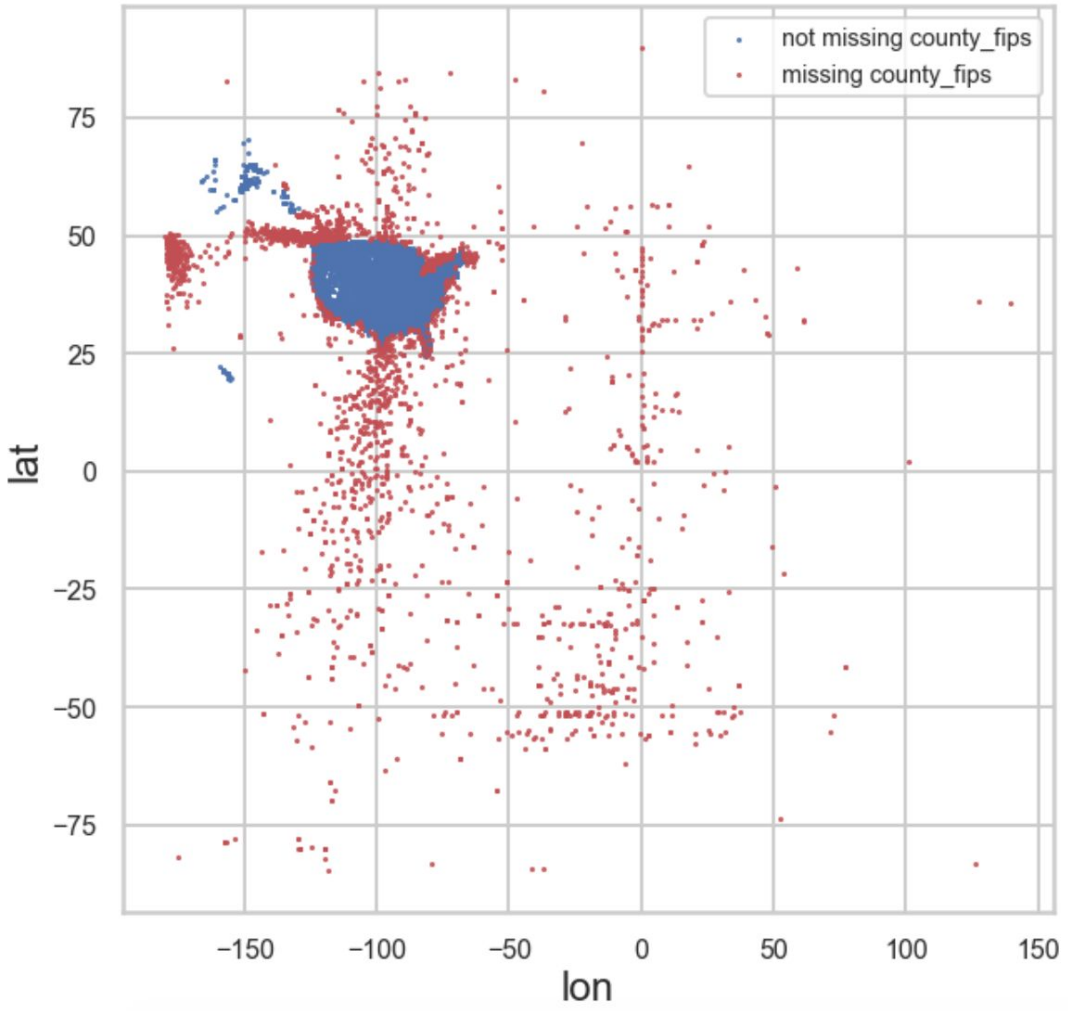
1.7M CAR POSTINGS FROM CRAIGSLIST

- Which manufacturer produces the fleet of cars with the best resale value?
- How would you go about detecting whether a listing is good or bad?
- Are there any other interesting insights or trends that are actionable?

URLs & GEOLOCATION FEATURES

Feature	% of NaN	Comment
url	0	URL
image_url	0	URL
city	0	GEO
lat	0	GEO
long	0	GEO
state_name	0	GEO
state_code	3.41	GEO
state_fips	3.41	GEO
county_name	3.41	GEO
county_fips	3.41	GEO

- Few missing values
- URLs are outdated, otherwise object detection could be used
- Reconstruct missing data using lat/long?



GEODATA

No categorical geoloc data

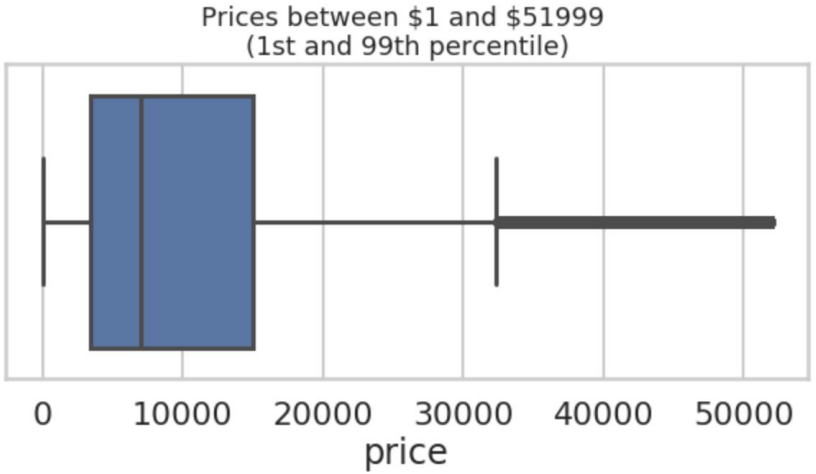
=

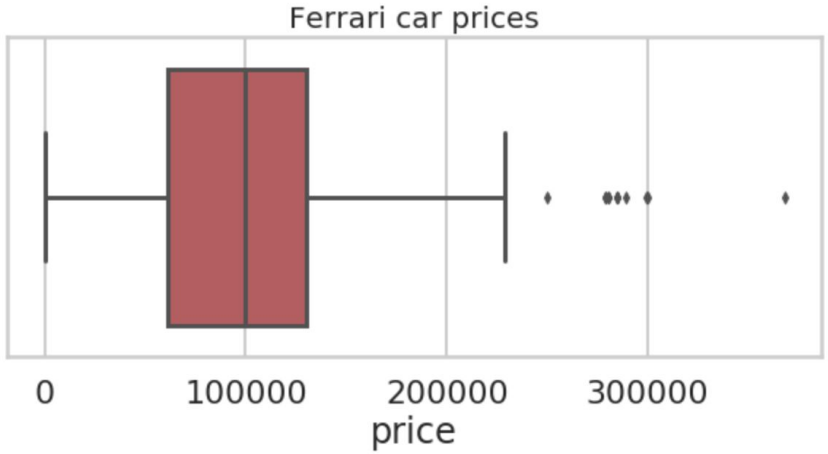
invalid lat/long

Feature	NaN, %	Comment	Suggestion
price	0	outliers	
year	0.37	outliers	
title_status	0.15	clean 93%	obligatory field
transmission	0.52	automatic 86%	obligatory field
fuel	0.6	gas 89%	obligatory field
weather	3.45	???	
make	4.05	100k uniques	select from a list
manufacturer	7.92	53 uniques	obligatory field
odometer	32.74	outliers	obligatory field
drive	38.41	4wd/fwd/rwd	
cylinders	40.12	8 uniques	
paint_color	40.37	12 uniques	
condition	40.67	6 uniques	obligatory field
type	40.8	13 uniques	obligatory field
vin	64.9	vehicle id	
size	65.23	4 uniques	obligatory field

MORE FEATURES

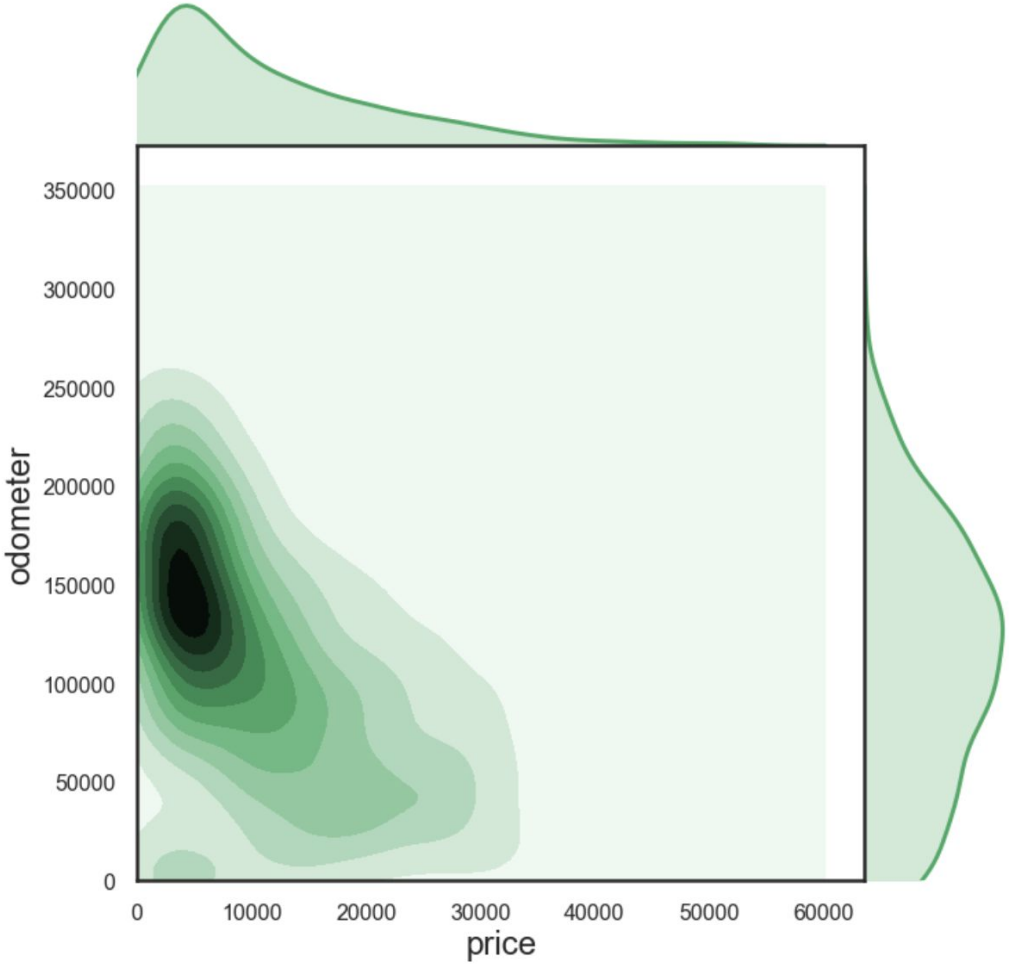
- Some features have a lot of NaNs, they should be made obligatory
- Price, year, odometer suffer from outliers





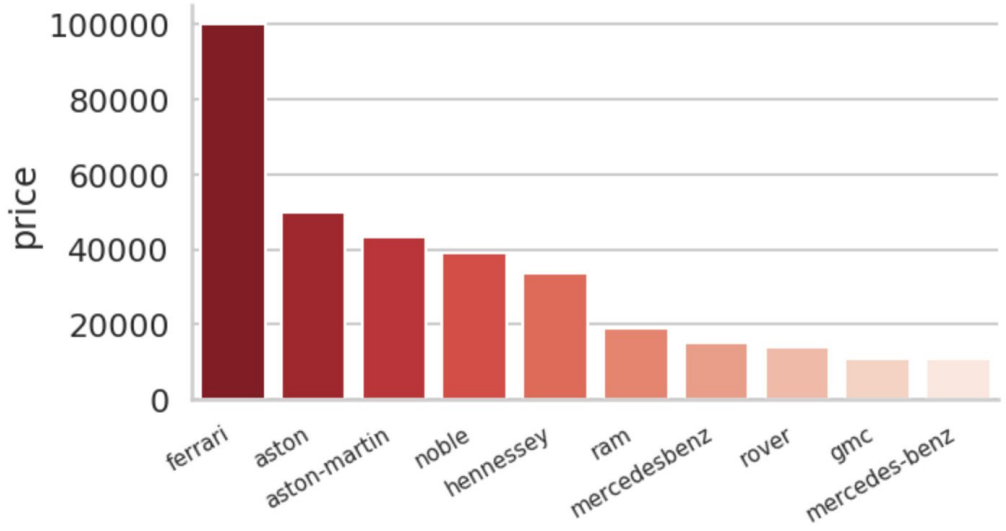
OUTLIERS IN PRICES

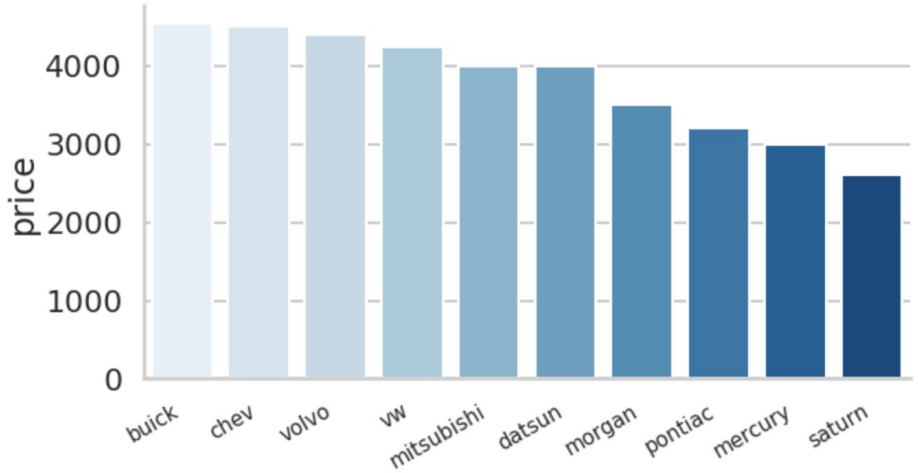
- Heavily skewed distribution
- Many entries with $1 price (fine to remove)
- Percentile based filtering removes high profile cars



ODOMETER VS PRICE

- Odometer defines price
- 33% NaN in Odometer
- Inputs with minimal price - $1
- Same with odometer - 0 miles





MEDIAN PRICE BY MANUFACTURER

- Highest: Ferrari with $100k
- Lowest: Saturn with $2.5K

PRICE DECAY WITH TIME

$$p = p_0 e^{\alpha \cdot \text{age}} \rightarrow \ln p = \ln p_0 + \alpha \cdot \text{age}$$

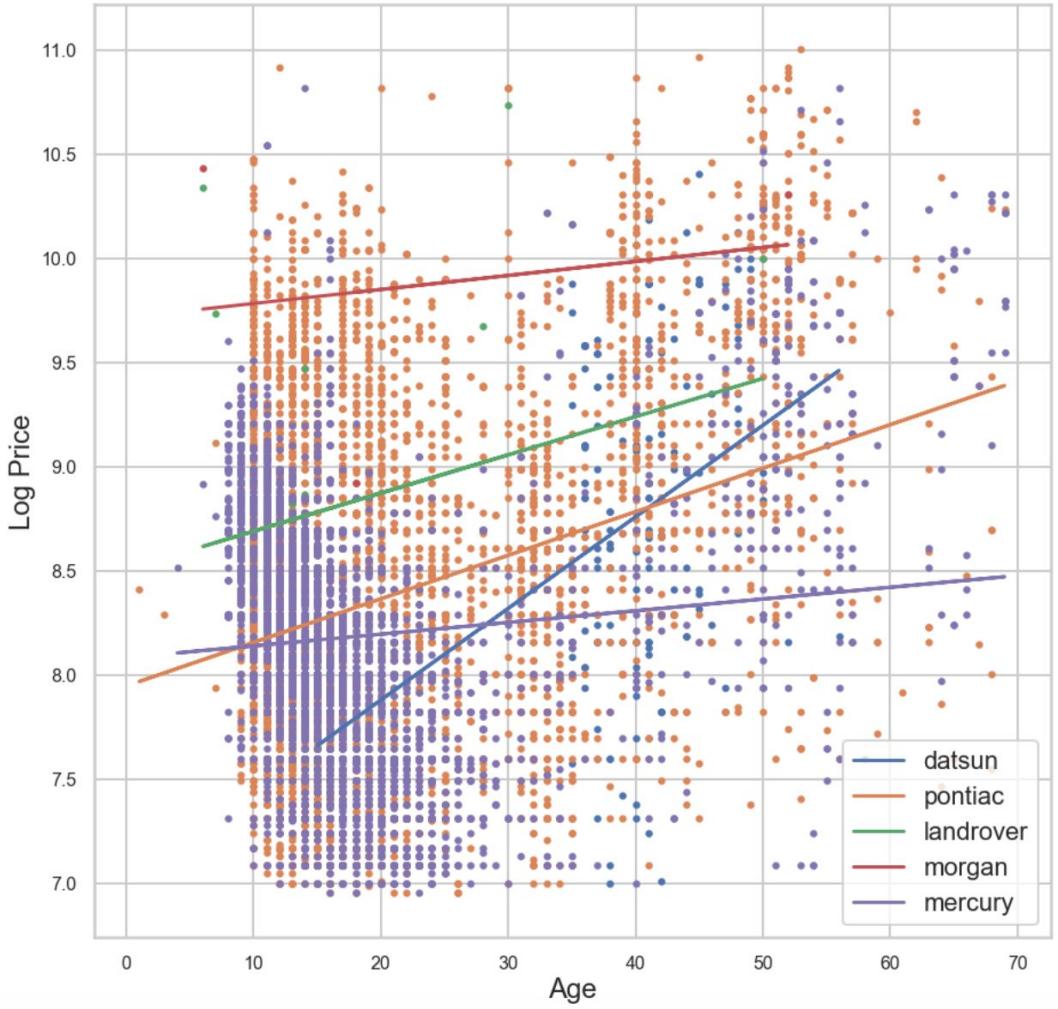
Linear regression with the original price as intercept and decay constant alpha as a coefficient

CUTOFFS

- Price between 1th and 99th percentile
- Mileage below 99th percentile
- Samples # >2

Brand	Alpha
datsun	0.043922
pontiac	0.020891
landrover	0.018323
morgan	0.006725
mercury	0.005622
chev	-0.005574
vw	-0.008545
chevy	-0.013428
fiat	-0.020576
ferrari	-0.0244

Datsun is an automobile brand owned by Nissan. By 1986 Nissan had phased out the Datsun name, but re-launched it in June 2013 as the brand for low-cost vehicles manufactured for emerging markets.

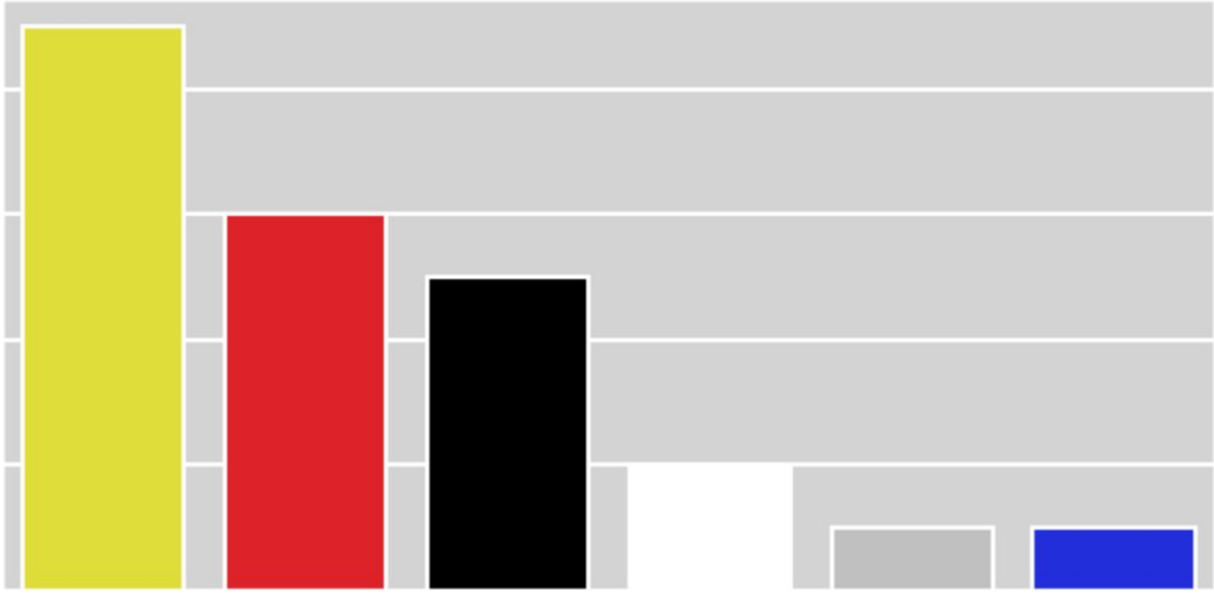


MORGAN

DATSUN

PONTIAC

BONUS
FERRARI PAINT COLORS*



*out of 24 cars with color specified

CONCLUSIONS

- Lots of missing data prevents building a reliable algo
- The problem can be solved at the data entry level by making some of the fields obligatory (odometer, year, condition)
- Geodata verification before submission
- Image analysis of the posting would be promising
- Remove outliers with high price might be costly, due to high potential losses