Used Car Data Challenge

John Henderson

September 25, 2014

Background

Recently, I started looking for a new [used] car. My search criteria:

- ~$6-8k
- Excellent reliability/low maintenance
- No signs of wheel well rust
- Manual transmission
- Working A/C (a first for me!)
- Not red
- Alloy wheels (preference)
- Sunroof (preference)

Hangup

How to quantify "reliability"?

- Consumer reports
- JD Power & Associates
- Common problems * materials/parts?

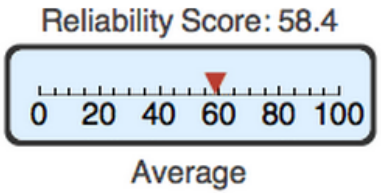
Cool!

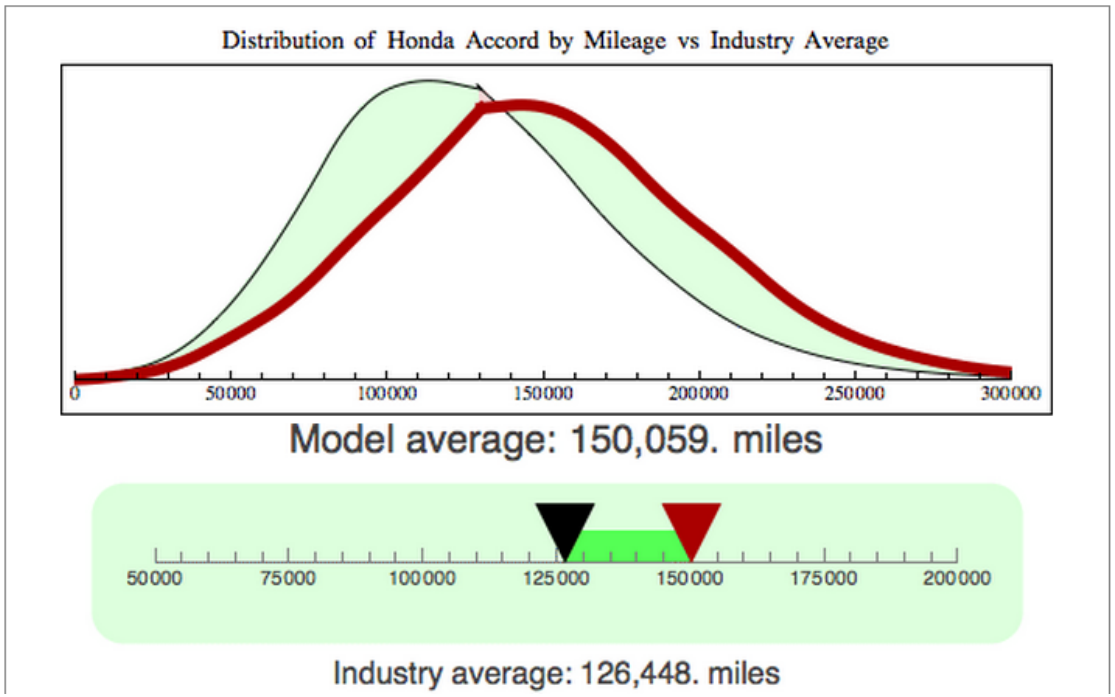
Stumbled on the Long-Term Quality Index:

- Agreements with numerous dealerships/retailers around the country
- Entities agree to report details on trade-ins
  - Year/make/model
  - Mileage
  - 0/1 for transmission or engine issues during inspection

Honda Accord

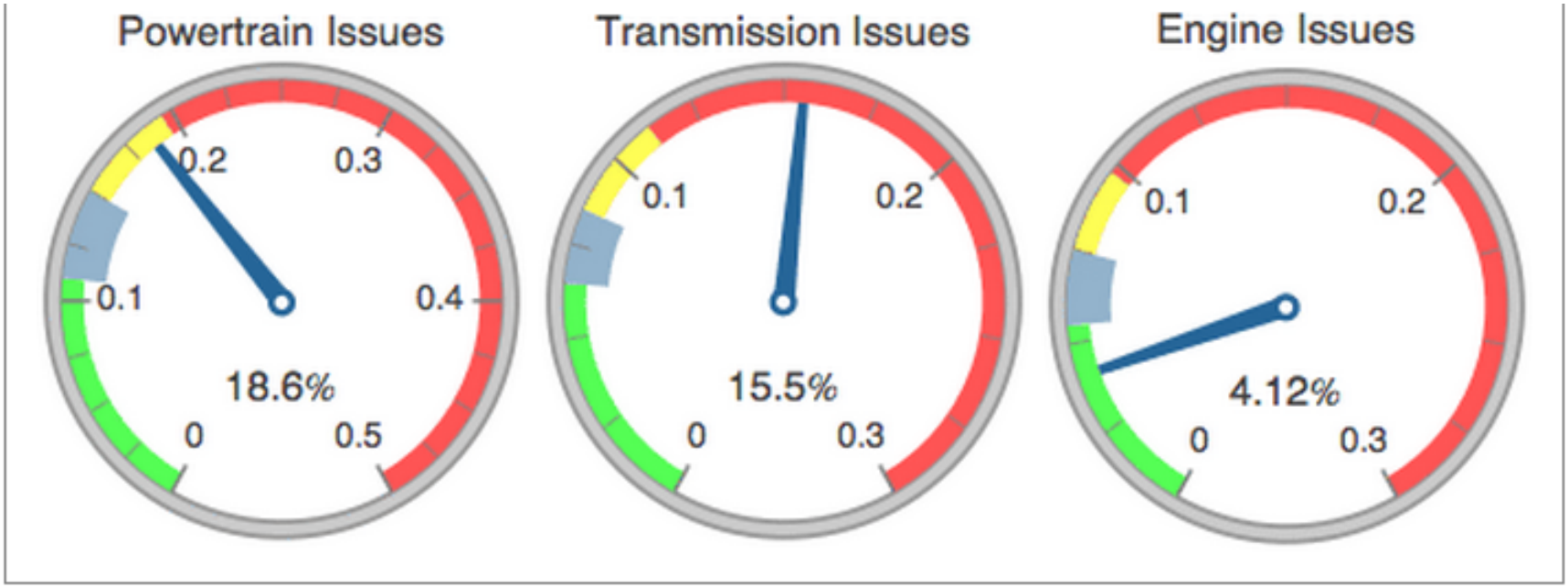
Mileage of make/model vs. distribution for whole dataset:





Honda Accord

Issues rate + "quality score":



- Need help reading this chart?

Sample Size: 14020

Contacted authors

Emailed the authors (because, why not?)

- Steve Lang, contributor to Yahoo! auto, GTA auto, etc.
- Nick Lariviere, Wolfram

Asked for sample data to play with and share with TCRUG

The data

Four makes/models from data set (~20k records):

- Honda Accord
- Toyota Avalon
- Chevy Cavalier
- MINI Cooper

Disbrituions of miles/issues for all cars in their data.

summary()

Removed 2014 cars (no good data) and mileage outliers.

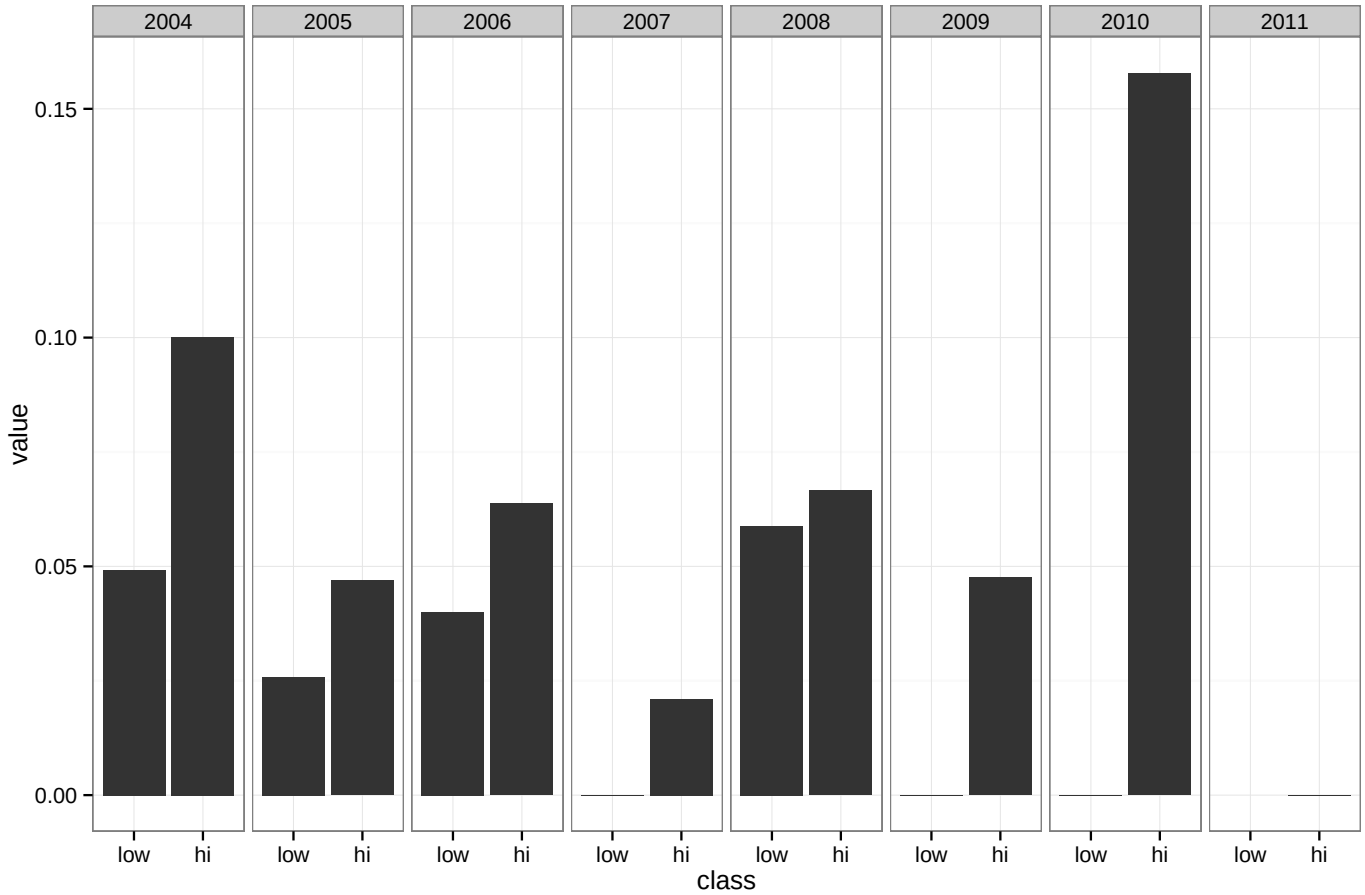
```
      year          miles           trans            engine       
 Min.   :1996   Min.   :   204   Min.   :0.0000   Min.   :0.00000  
 1st Qu.:2000   1st Qu.:100174   1st Qu.:0.0000   1st Qu.:0.00000  
 Median :2002   Median :137076   Median :0.0000   Median :0.00000  
 Mean   :2002   Mean   :139157   Mean   :0.1265   Mean   :0.05189  
 3rd Qu.:2005   3rd Qu.:175286   3rd Qu.:0.0000   3rd Qu.:0.00000  
 Max.   :2012   Max.   :394991   Max.   :1.0000   Max.   :1.00000  
```

Getting started

- Data is on the TCRUG github site
- Uploaded some starter code
  - Read in the data
  - Munge some ugly database output with distributions
- Some questions to prime your brain

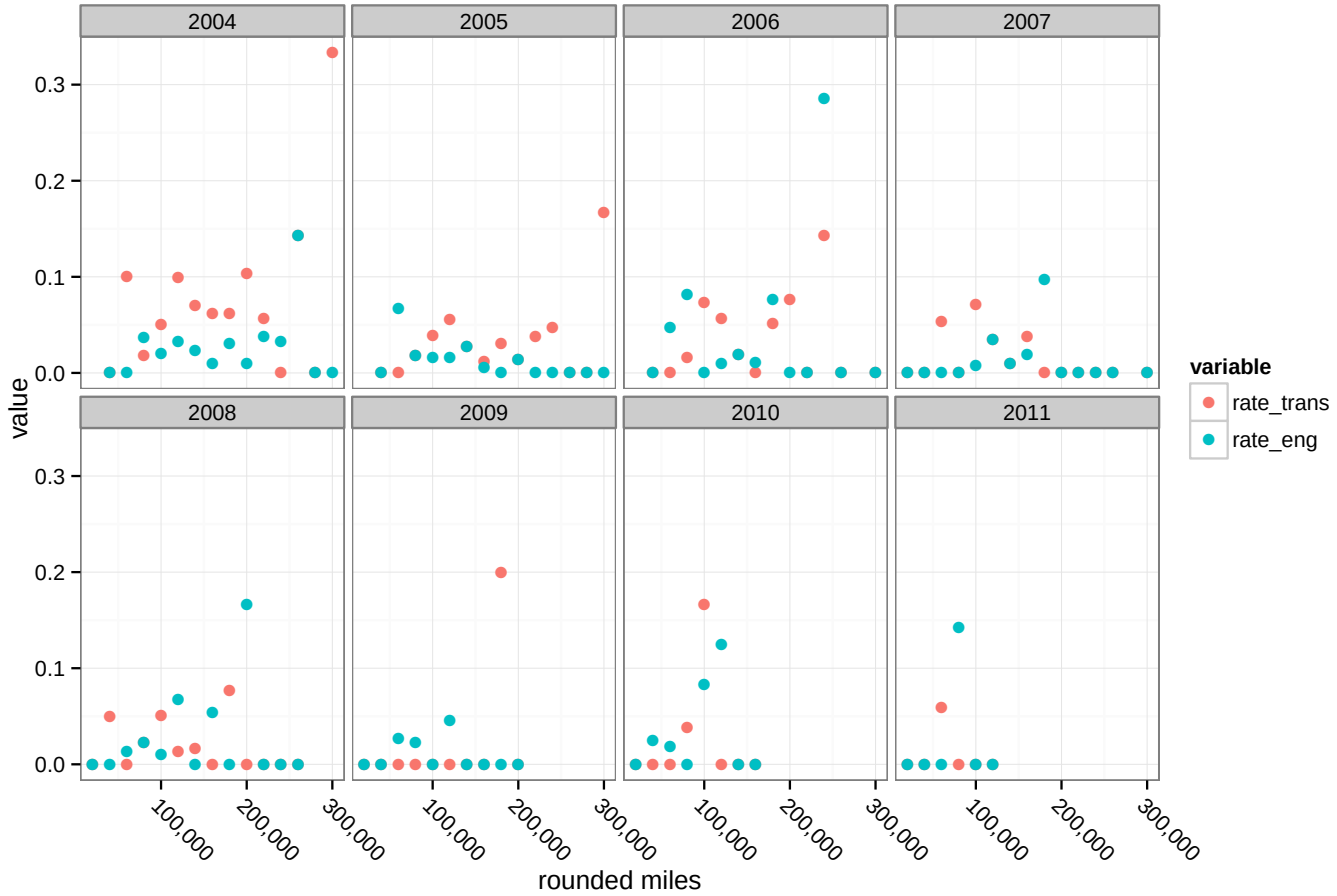
Some plots for ideas

Lower vs. higher mileage cars?



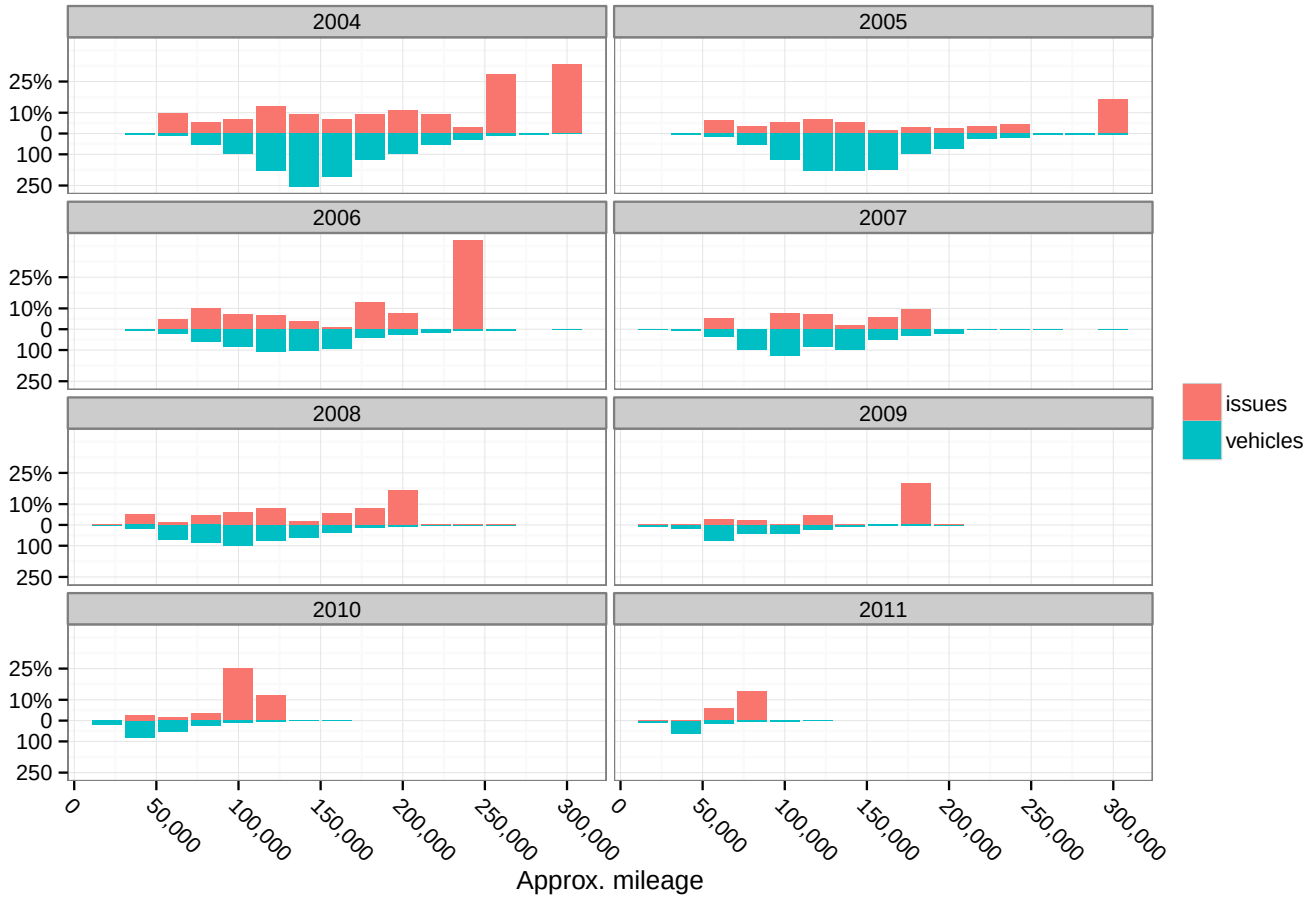
Some plots for ideas

Any lemon years?



Some plots for ideas

How to capture robustness of issues rate (via `n`)?



Next month

Come with your ideas, visualizations, insights, etc.!

- Cool plots
- Differences between various makes/models
- How to compare used cars for reliability
- New derived metrics
- Better visualizations for the LTQI site

I'll be providing feedback to Steve and Nick for v2 of their site!