

# Midterm Proposal

Group 7 Chenxi Di, Nan Tang, Ying Wang, Yining Wang  
Group leader: Nan Tang

## Section 1

### Data Source

We obtained our data through Kaggle: <https://www.kaggle.com/jpayne/852k-used-car-listings>.

### Description of Data

Our dataset includes all available used car listed on TrueCar.com September 24th 2017. It has 1.2M listings of used cars scraped from TrueCar.com. Each row represents one used car listing.

### Variables

- Price: numerical data indicating price of used car.
- Year: each vehicle has a consistent model year, in which it records the starting year that the car is legally go on sale. All cars in the dataset have model year ranges from 1997 to 2017.
- Mileage: numerical data deriving from car's odometer reading.
- City: current location of the cars. (2553 levels)
- State: the state where these cars placed in. (51 levels)
- Vin: unique identification of each car, serving as a vehicle's fingerprint.
- Make: the manufacturer / brand of the vehicle. (58 levels)
- Model: the model type of the car, cars of different brand may use same model. (2736 levels)

## Section 2

### Goals

- Estimate the availability of used cars given certain conditions (price, mileage, year). I.e. derive the density functions of car availability on price, mileage and year from this dataset. The dataset includes only one-day's car information, thereby, our job is to apply resampling method to this 'big sample'.
- Build a model that can determine the price of used cars based on all given data.

Factors consist of price, year, mileage, make, and model, considering these factors are major concerns of car pricing. One of the purpose to build such model turns to give the clients a uniform criterion when choosing used cars.

E.g. After input specified values such as mileage, year, model type, make and the price that trader provides, the model will come up with reasonable range of pricing whether the trader's pricing is reasonable.

- Find the possible relation between the scale of the cities(determined by population) and the sales prices for used cars.