MA678 Midterm Proposal

Yuli Jin

2021/11/4

Data scource:

https://www.kaggle.com/lepchenkov/usedcarscatalog
The dataset I plan to use in this project is the used car catalog dataset. The dataset was scraped in Belarus on the 2nd of December 2019. The dataset contains listed price, brand information and specification of used car.

Personal Statement:

My career goal is to be a data analyst in finance and business field. Building an effective model to predicting the price of house cars and other assets is common in many asset appraisal firms. Also, having a valid interpretation of which factor contributes to the price helps people better understand their asset. Compared to house dataset, the used car dataset contains multilevel variable which complies with the project requirement (at least 10 groups). Preliminary, I plan to use log(price) as the response variable and regard year, odometer value and car body type as explanatory variable. Some brands varibale may be applied to conduct multilevel model.

The questions I'm trying to answer are listed below:

How these variables affect the used car price?
Whether the residuals of the model conform with assumptions?

The proposed timeline of work is listed below:

EDA & data processing: before 11/14/2021
Modeling and Validation: before 11/21/2021
Writing : before 11/28/2021