4 Implementation/Design

This chapter will describe the process of implementing the system. The implementation was divided into five parts titled Data Set, Data Cleaning and Normalization, Machine Learning Algorithms, Measurements, and Inference. Each of these parts are explained in their own sections as part of this chapter and are shown in the UML diagram below (see figure 1). The high level component of the UML diagram without a dedicated section of this chapter, Simulated Aging, is detailed in the measurement section. The entire implementation was written in Python3 in the PyCharm ide. The libraries utilized are pandas, sklearn (sci-kit learn), NumPy, re (regular expressions), matplotlib, and seaborn. (See Appendix C for the entire source code)

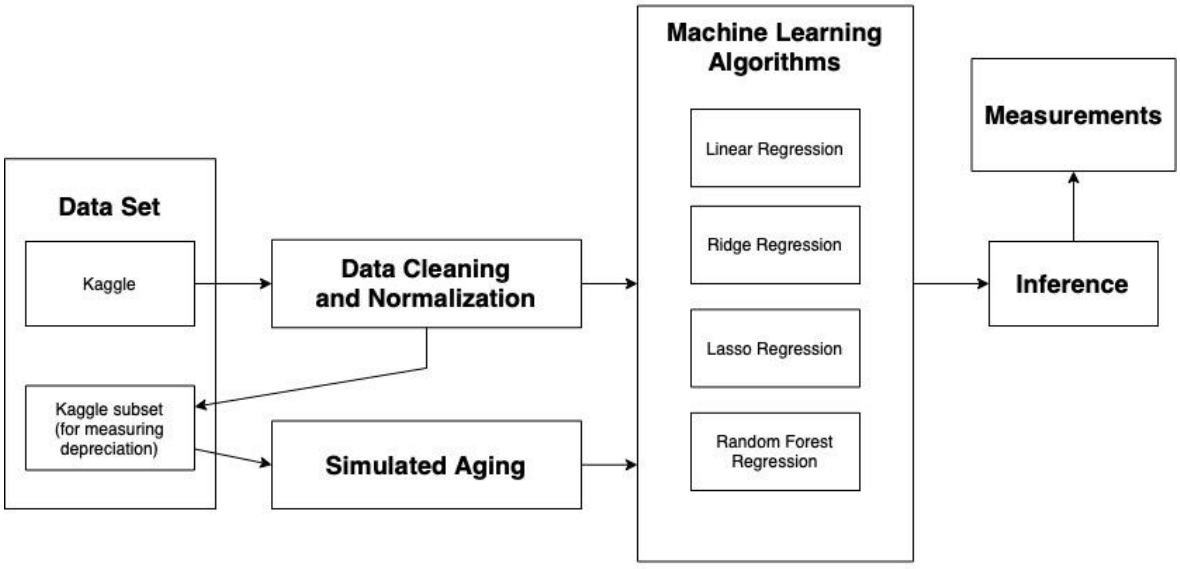

Figure 1: UML Component Diagram, Implementation Overview

4.1 Data Sets

4.1.1 Kaggle Dataset

The dataset was sourced from Kaggle and includes 122,144 car listings from the years 2018, 2019, and 2020 from all areas in the United States. It is available publicly. It includes all types of road-going consumer