1. Introduction

The dataset is about a survey of customers flying within the United States. It has 129889 rows with 28 variables which includes data for three months(January,February,March). The main goal of this project is to increase customer satisfaction for the airlines.

The dependent variable in this dataset is the Customer Satisfaction whereas there are 27 independent variables. These are in various subcategories such as personal attributes (age, gender, price sensitivity,…), flight planned status (class, date, Airline, Origin City, …) and how the flight went (departure delay, flight time,...)

2. Data cleanse/munge/preparation

The cleaning of data started with finding out the NA's available in the dataset. It was found that the NA's were available in three columns: Departure Delay, Arrival Delay and the Flight time. The three steps we followed are:

- Replaced the NA's in departure delay, arrival delay and the flight time with a large value when the flight was canceled. (Flight.cancelled =”Yes”). This means that because we had a larger delay, the flight was cancelled.
- Deleted the 337 rows which had NA's in the departure delay, arrival delay and the flight time when the flight was not cancelled. Because, in a large dataset with 129889 rows, 337 has a percentage of 0.2.
- There are three satisfaction values with a discrepancy which is deleted. Therefore there are 340 rows deleted in total.