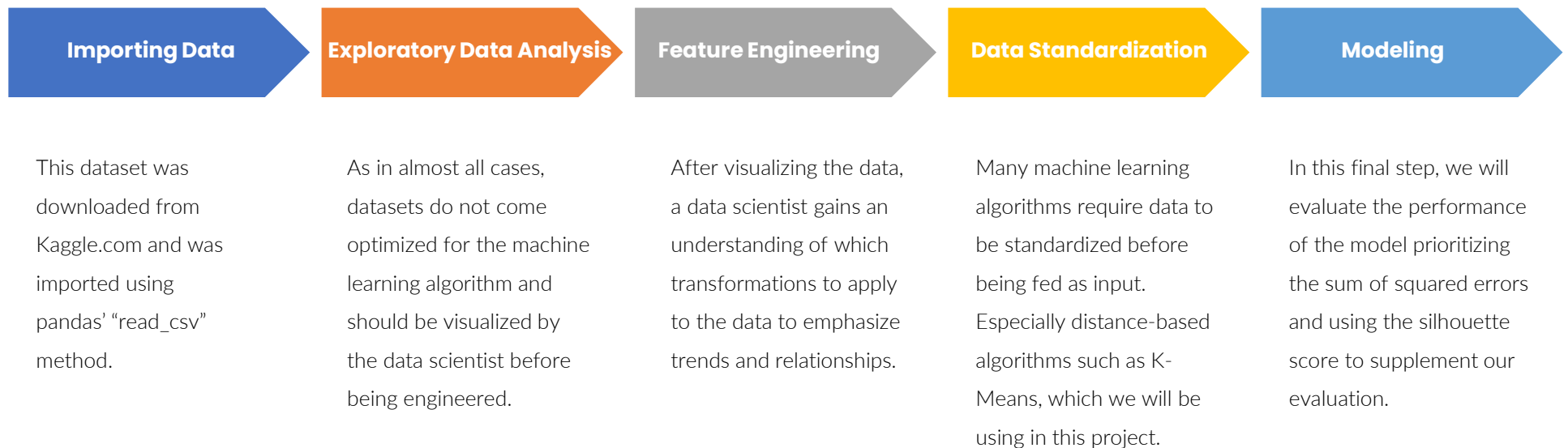# Customer Clustering

Jonathan Lampkin
January 2022

Springboard Capstone 3

In this project we will focus on clustering customers by similarity. This step precedes the handling of individual clusters for various purposes.

Project Walkthrough

Importing Data

Exploratory Data Analysis

Feature Engineering

Data Standardization

Modeling

This dataset was downloaded from Kaggle.com and was imported using pandas' "read_csv" method.

As in almost all cases, datasets do not come optimized for the machine learning algorithm and should be visualized by the data scientist before being engineered.

After visualizing the data, a data scientist gains an understanding of which transformations to apply to the data to emphasize trends and relationships.

Many machine learning algorithms require data to be standardized before being fed as input. Especially distance-based algorithms such as K-Means, which we will be using in this project.

In this final step, we will evaluate the performance of the model prioritizing the sum of squared errors and using the silhouette score to supplement our evaluation.

Quick View of the Dataset

	AGE	FLIGHT_COUNT	AVG_INTERVAL	MAX_INTERVAL
0	39.0	1.945910	109.000000	274
1	60.0	2.944439	33.666667	126
2	39.0	2.995732	33.368421	124
3	37.0	2.397895	37.500000	80
4	29.0	1.386294	53.000000	132

Age	Flight Count	Average Interval	Max Interval
Age of the individual	Total Flight Count of the individual	Average number of days an individual does not fly	Maximum number of days an individual has not flown

|	count	mean	std	min	25%	50%	75%	max
AGE	55000.0	42.701400	9.773425	6.000000	35.000000	42.000000	48.000000	110.000000
FLIGHT_COUNT	55000.0	2.159785	0.897792	0.693147	1.386294	2.079442	2.772589	5.361292
AVG_INTERVAL	55000.0	62.979631	64.506386	0.000000	25.166667	44.600000	77.500000	714.000000
MAX_INTERVAL	55000.0	172.201745	117.382011	0.000000	90.000000	149.000000	230.000000	719.000000

Feature Distributions
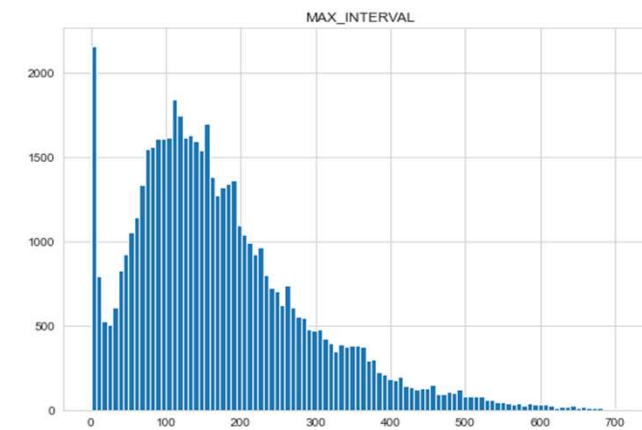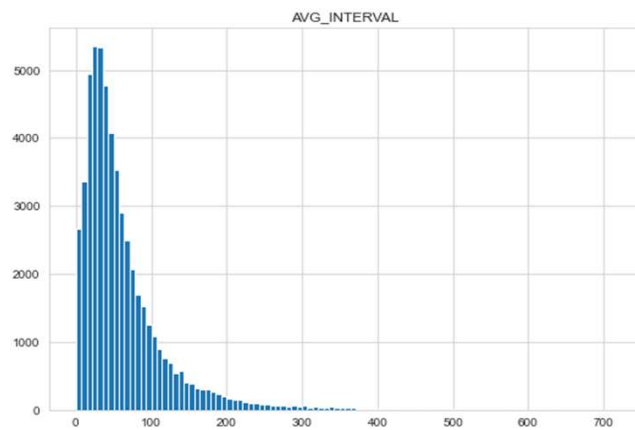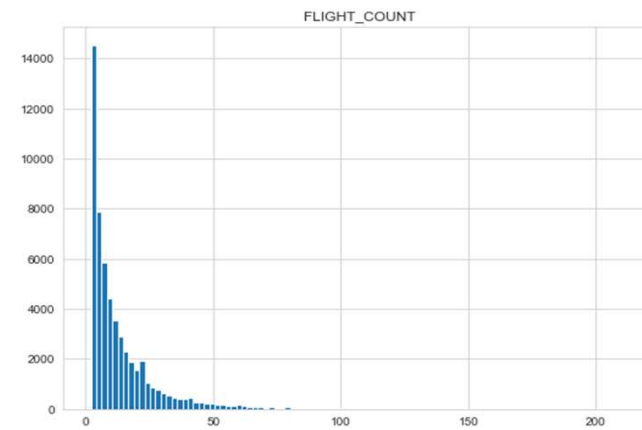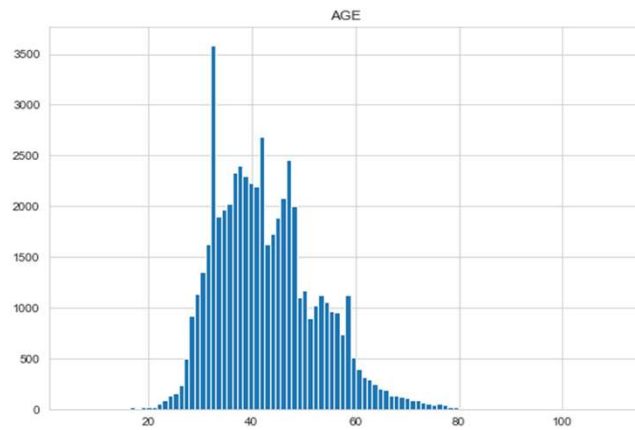
Pairplot
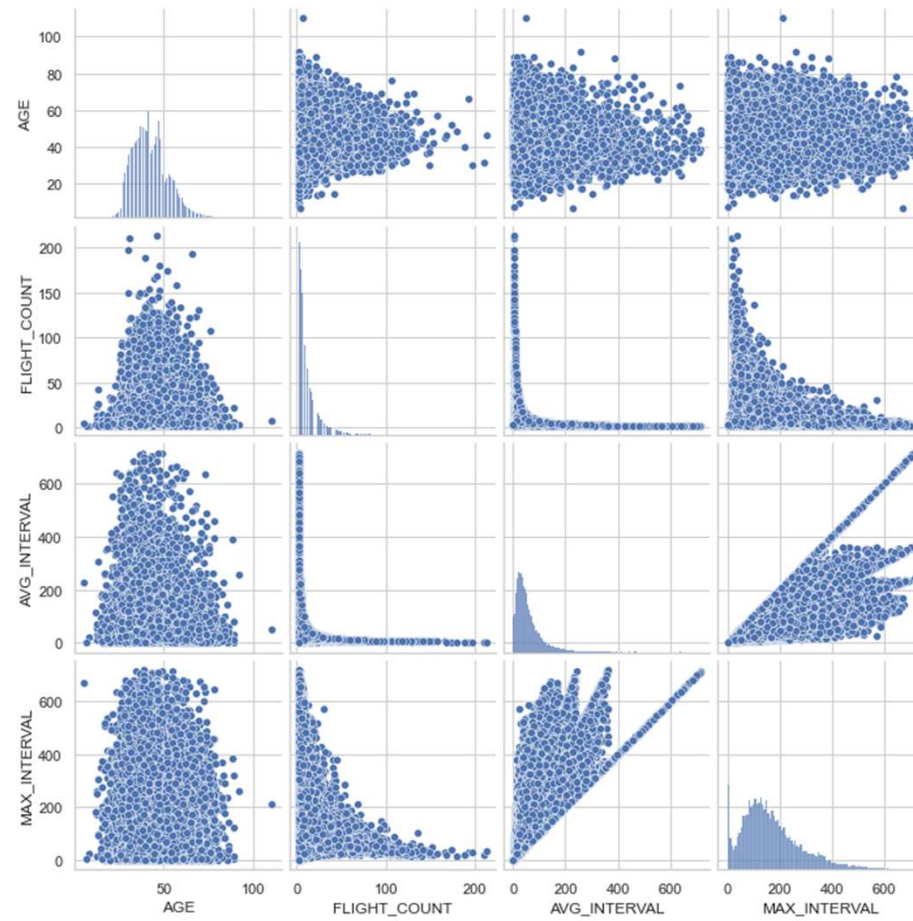
Feature Correlations
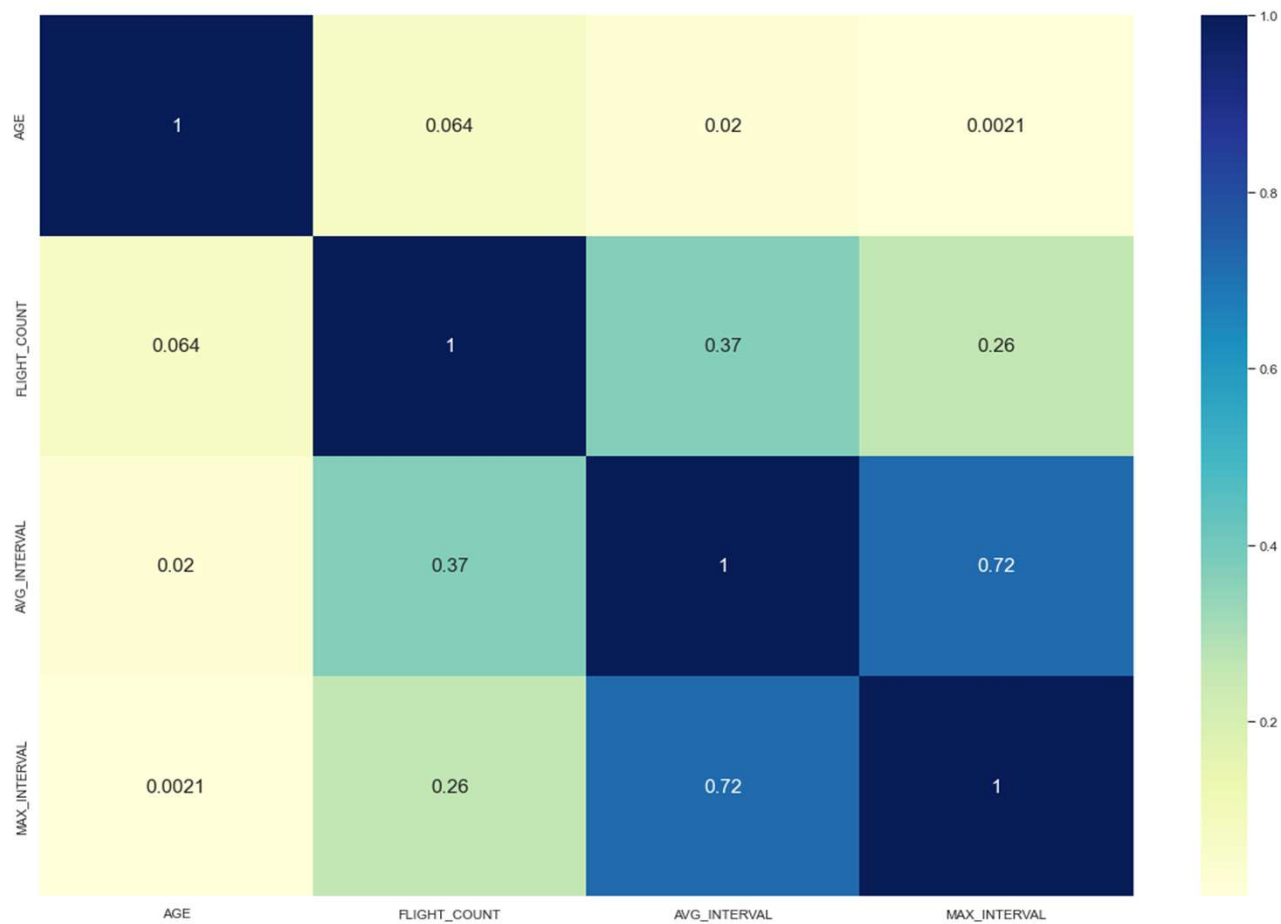
Most Frequent Flyers
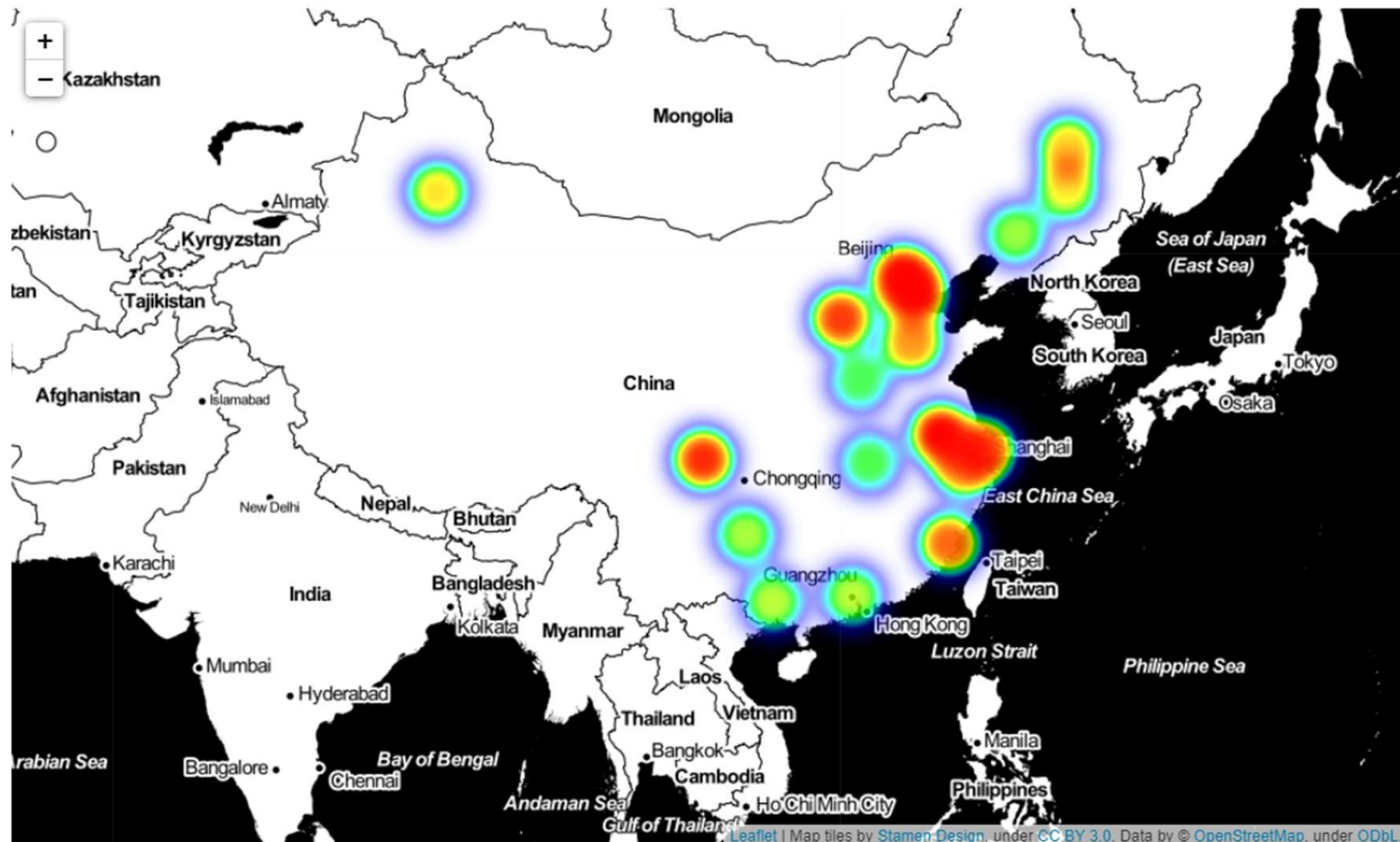
Model Performance Evaluation

SS for $2 \leq K \leq 50$

Silhouette for $2 \leq K \leq 50$
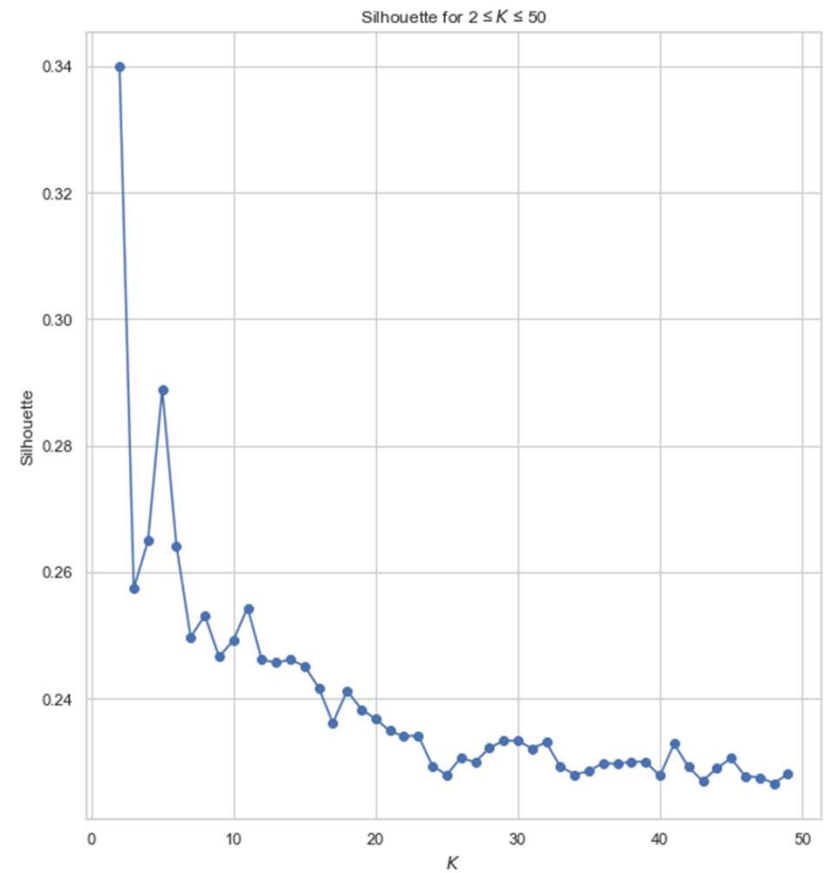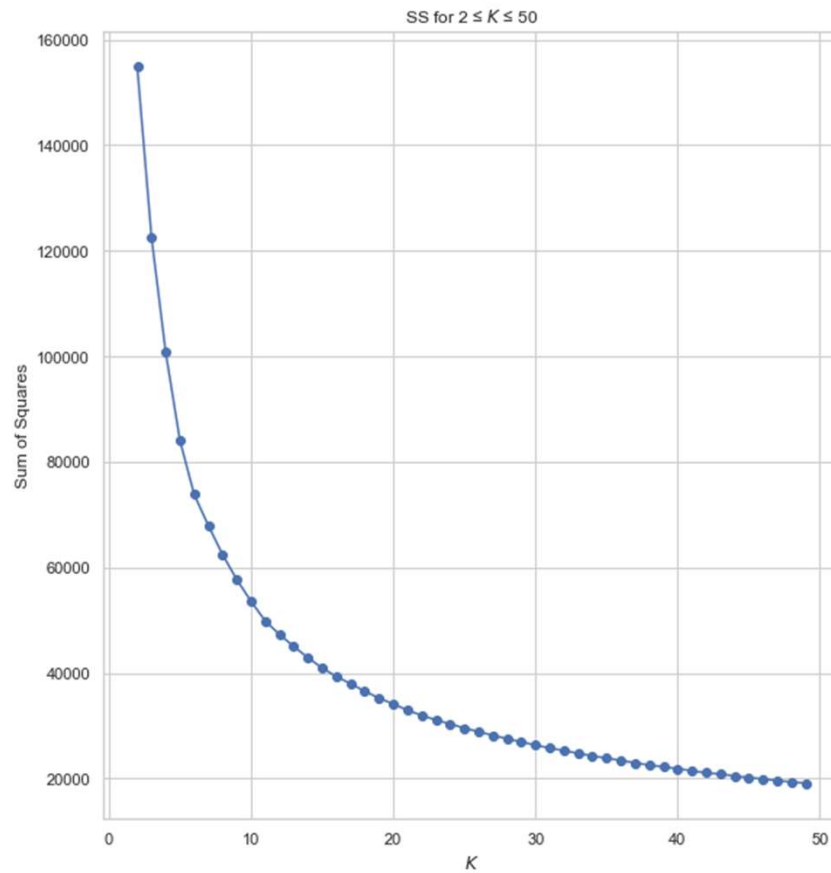
Resulting Clusters!
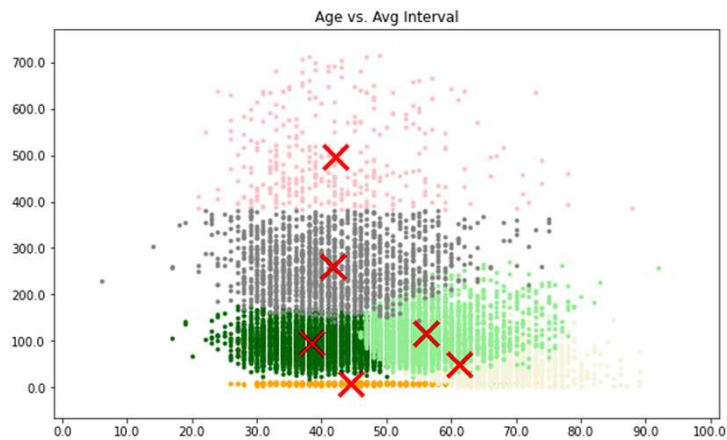
K-Means created some awesome clusters that visibly separated individuals in a way that can be understood.