Airline Customer Clustering – Springboard Capstone 3

Introduction

This file documents my third and final project under the Springboard Data Science Career Track Curriculum. For my final project I chose to deepen my experience, understanding, and familiarity with the unsupervised learning method: Clustering. We will use Scikit-learn's K-Means Clustering Algorithm to perform the task of clustering airline travelers into groups dependent on the features fed to the algorithm. This dataset didn't come with an explanation for its features and thus most of them were dropped due to an inability to understand what they represented. Below is an image representing the first five rows of the dataset and the features that will be used in this project.

	GENDER	WORK_CITY	WORK_PROVINCE	WORK_COUNTRY	AGE	FLIGHT_COUNT	AVG_INTERVAL	MAX_INTERVAL
0	male	shanghaishi	shanghai	cn	38.0	2	186.000000	186
1	male	guangzhou	guangdong	cn	63.0	22	33.857143	368
2	male	guangzhou	guangdong	cn	53.0	14	46.307692	151
3	male	beijingshi	beijingshi	cn	43.0	5	101.500000	393
4	female	zhengzhou	henan	cn	42.0	19	35.833333	125

Problem Statement

Using the interpretable and meaningful features of the dataset: Age, Average Interval, Max Interval and Flight Count. We will cluster individuals/observations based on their travel frequency, total amount of flights, and their age. Success for this task will be measured by visibly separate clusters, and three metrics used to evaluate unsupervised clustering algorithms: Elbow Law, Sum of Squared Errors, and Silhouette Score.

Methodology

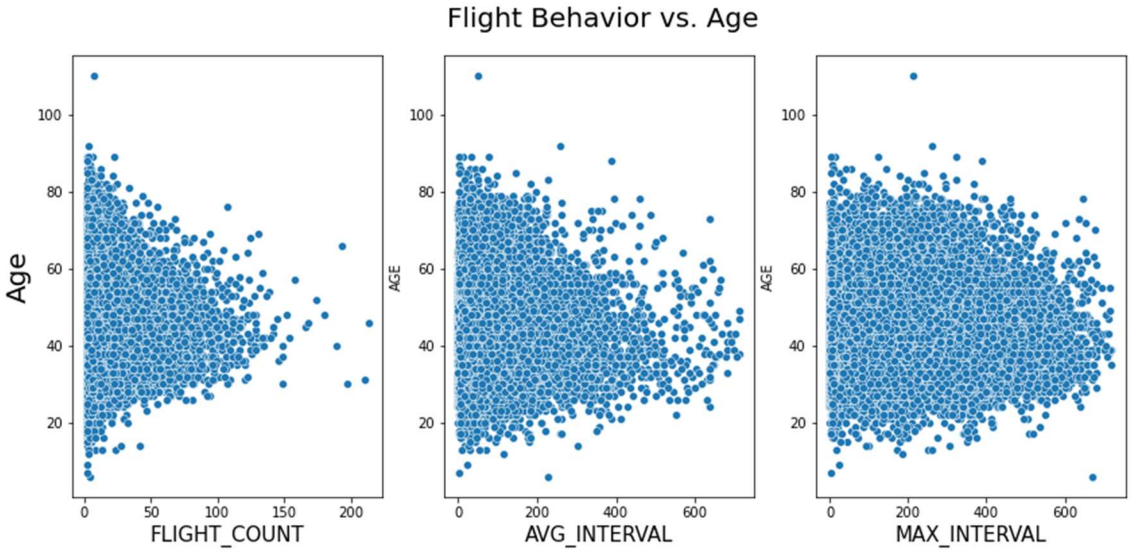
Exploratory Data Analysis

As with every project, we start by seeking to understand the dataset's features and thinking intuitively about which features may relate to our target objective. After dropping most of the features that could not be understood or did not relate to our objective, we explored summary statistics for the remaining features, and evaluated null values. Many of the numeric features had outlier variables and we chose to fill null values with the median.

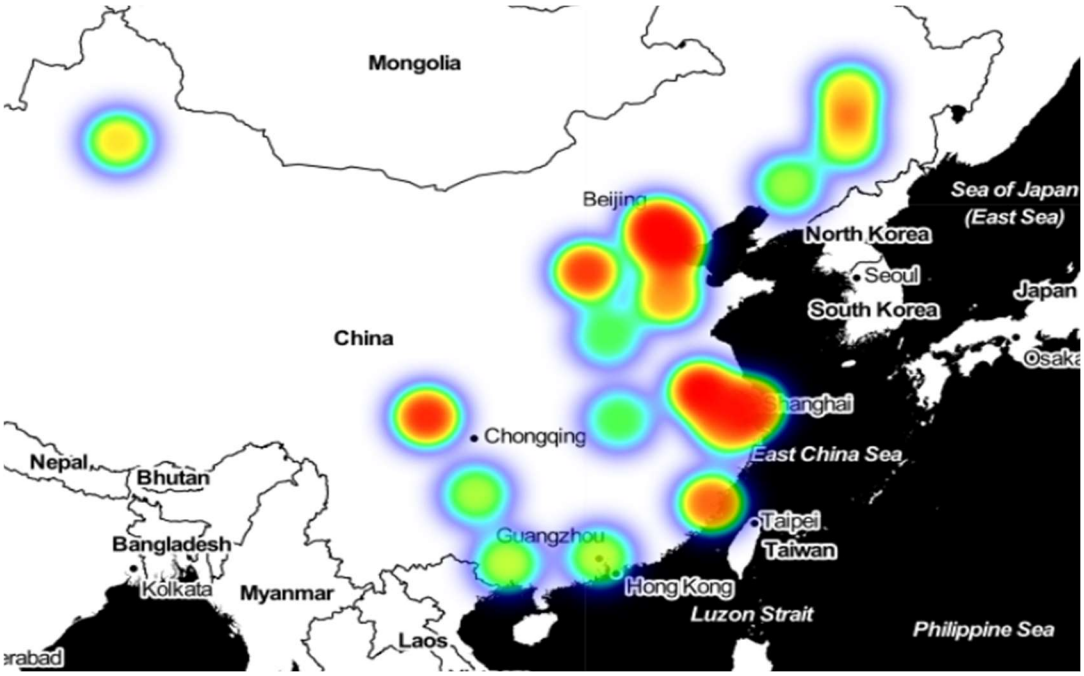
We explored feature distributions using a histogram and explored linear correlations using Seaborn's heatmap along with Pandas' correlation method. We were particularly interested in finding a relationship between demographic features (Gender and Age) and flight habit features (Flight Count, Average/Max Interval). The Gender feature did not have significant correlation with features that described flight behavior, and Age did not have a linear relationship with flight behavior.

We dug deeper to explore the possibility of a non-linear relationship existing between Age and flight habit features by using Seaborn's pairplot method on the dataset. This visualization revealed a non-linear relationship between Age and Max Interval, Average Interval, and Flight

Count. The image below shows the non-linear relationship Age has with other flight habit features.



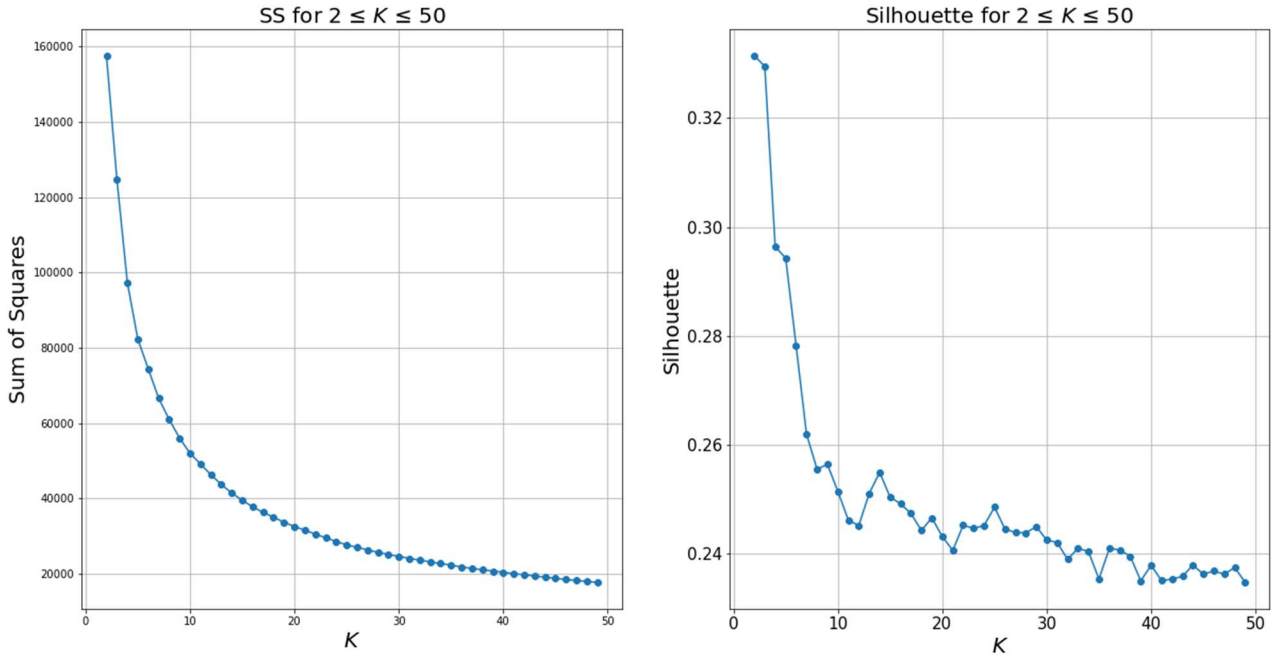
Our last exploration of the data was to use the work province feature to evaluate where our most frequent flyers were from. The data in this feature was extremely dirty and required quite a bit of mapping. Once cleaned, we used Google's Geocode API to convert the provinces to latitudinal and longitudinal data. These coordinates were passed to Folium, a mapping library along with Average Interval to create this beautiful display showing where our most frequent flyers come from.



It appears that the areas surrounding Beijing and Shanghai have the most frequent flyers.

Modeling

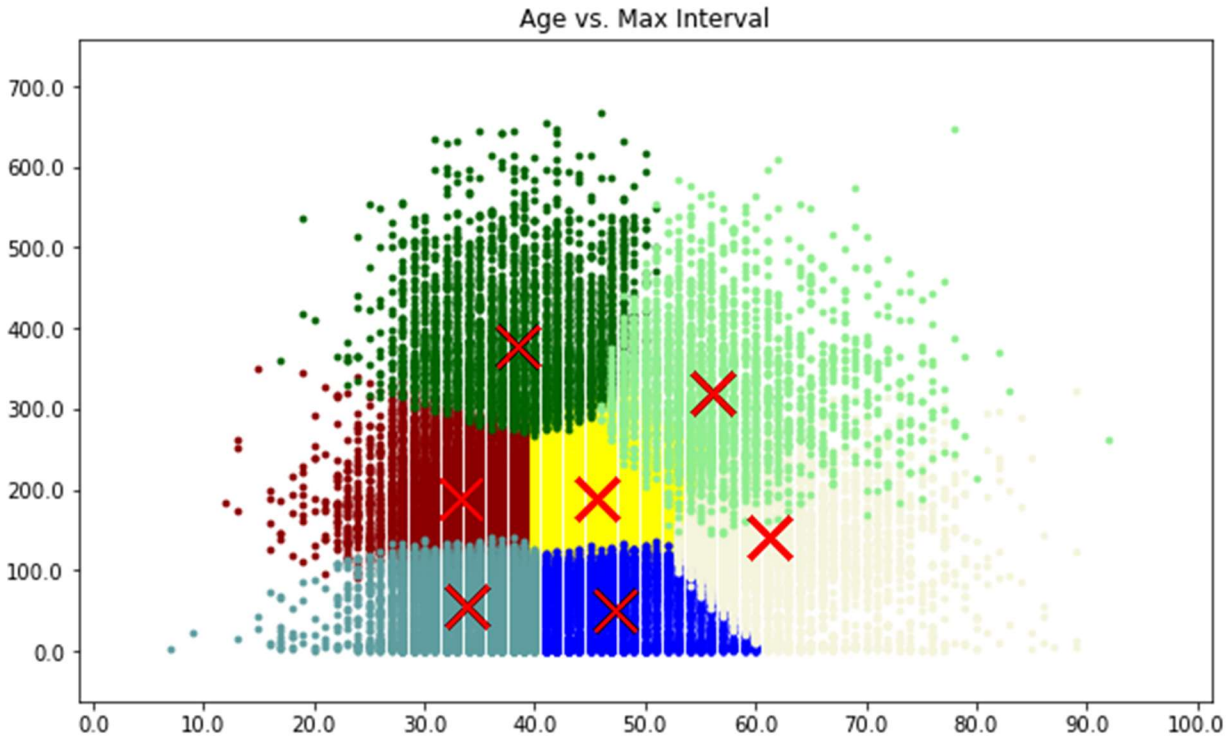
The first step of modeling our data was to apply the Z score function to it - as most algorithms, especially distance-based ones, prefer standardized data. We used the K-Means algorithm to cluster this dataset and thus, went through the process of deciding K (the ideal number of clusters). This is a fun process that requires data scientists to call on their artistic side. This is because K is an arbitrarily chosen parameter when using the Elbow Law, which is the most common and practical method for deciding K. We used the Silhouette Score metric, which measures the structure of K clusters (or goodness of fit), to supplement our decision for K. Below is pictured two graphs, the left one showing K (Number of Clusters) vs. the Sum of Squared Error, and on the right is the Silhouette Score at each interval of K. The Silhouette Score ranges from [-1,1], with a higher score representing a better structure.

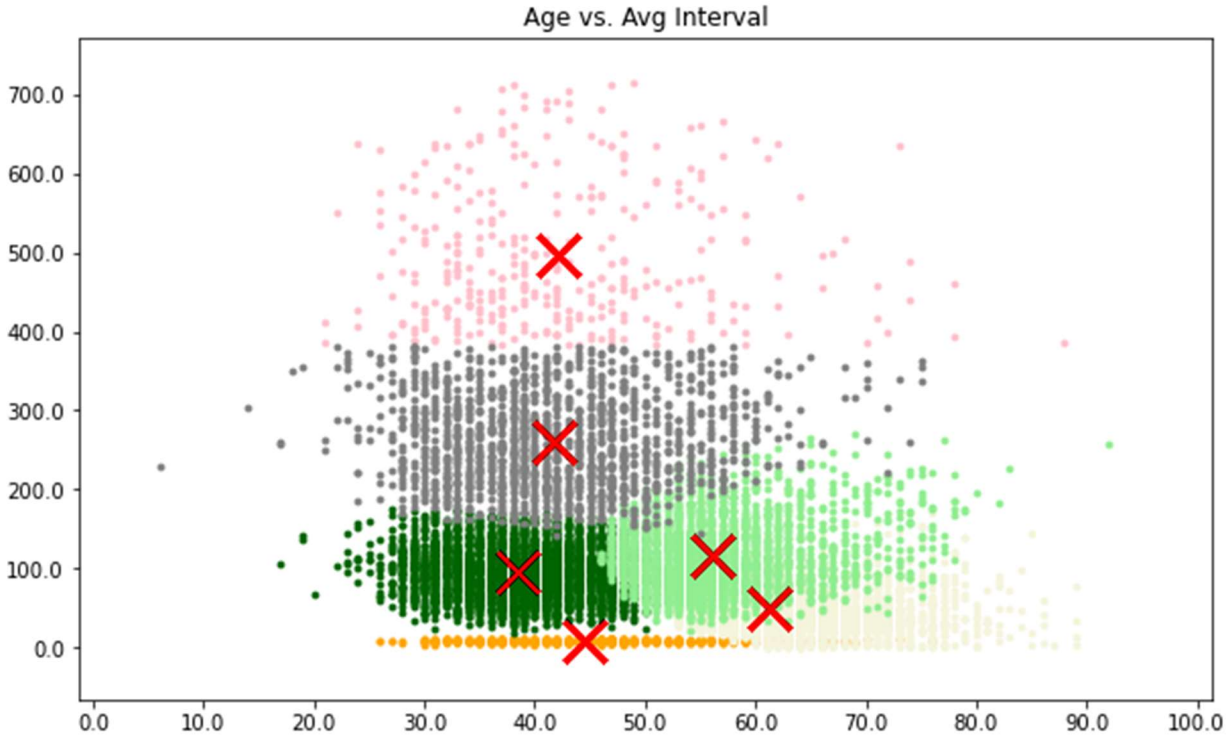


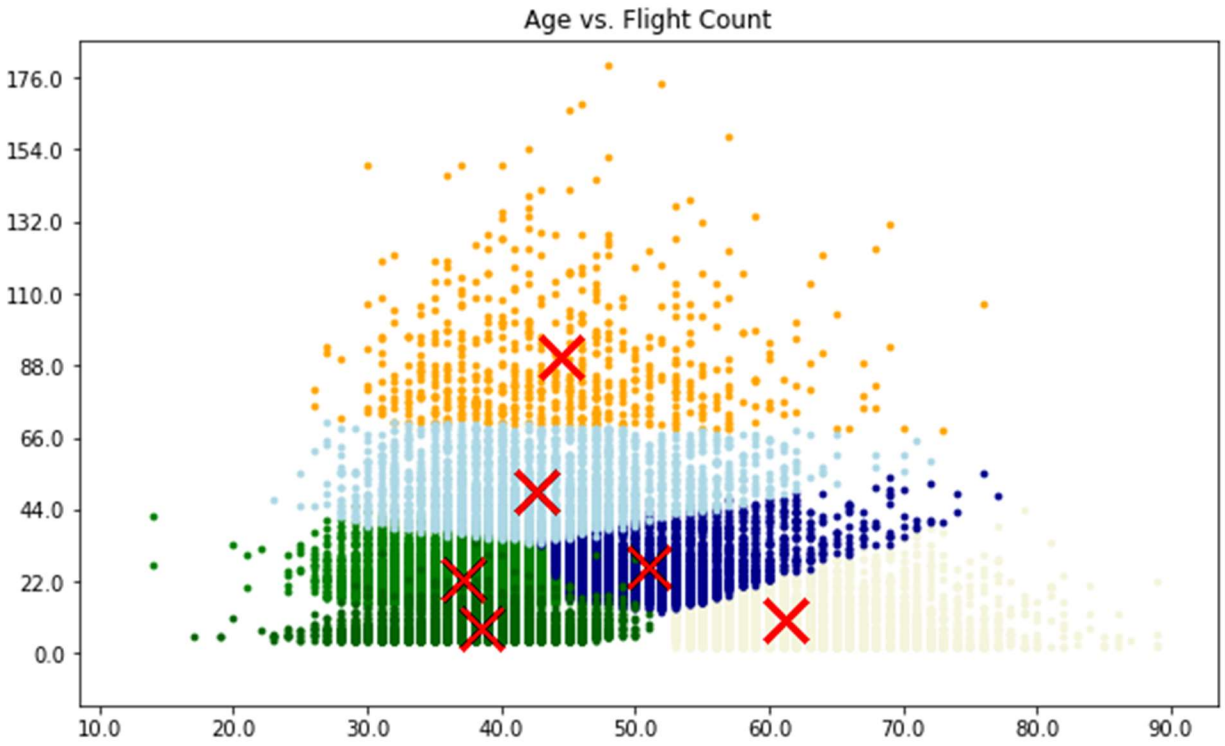
Looking for the pivotal point in the graph on the left combined with the highest point in the graph on the right, we decided 14 was a suitable number of clusters (K).

Results

At last, we can visualize the resulting clusters creating by the K-Means algorithm. Below are a few clusters generated by the K-Means algorithm to group individuals based on similarity.







Additional Notes

K-Means produced some great clusters that can be useful for classifying new individuals into groups with similar flight habits. In the future, I hope to work with datasets that have more demographic features which can improve the potential of the clusters. Including location as a feature could have yielding some interesting clusters as well.