Dataset 2

One of the challenges of the second dataset was to properly parse the dataset. The python packagee *arff* was not able to read the dataset, because one of the airline carriers was marked as "PA (1)". For this reason, we read line by line and parsed the dataset with a for loop. The list was transformed in a panda's data frame. However, the columns of the data frame were identified as object type, which is a non-suitable data type to apply the desirable functions. For this reason, we wrote the data frame in a csv file and read again the csv setting proper data type. We used pandas *scatter_matrix* function to plot the distribution of the features. Moreover, we printed the dataset size and the features as requested by the exercise.

For the second task of the exercise ("Describe the average delays per airport/airline"), we used pandas and *groupby* function to investigate the average delay per airline and per airport (destination airport and origin airport). We noticed that the column ArrDelay has positive and negative values. Therefore, we filtered flights that arrived without delay or earlier.

The third task ("Identify and report the most prominent rules of association between delays and point of origin AND/OR point of arrival.") was dealt with a similar manner. We grouped by both origin and destination airport and used *count* function to investigate the routes with the higher number of delayed flights. Moreover, we counted the delayed flights by origin and destination airport.

For the forth task ("Try to predict the delay given all other features and report the appropriate performance on cross-validation."), we detected and removed outlier instances using the function *stats.zscore*. 1857 flights with extreme delays were removed. Moreover, we plotted the outlier detection in a scatter plot. Afterwards, we converted the categorical columns in multiple columns with zero and one using the function *get_dummies* from pandas package. The dataset was split in training set and test set using the function *train_test_split* from sklearn package. We performed model selection by trying four different models (linear regression, ridge regression, elastic net, K neighbors regressor) and 10-fold cross validation. Ridge regression showed the best performance. We used grid search to tune the ridge regression hyperparameters and trained the model. We evaluated the performance on the test set using as performance criteria the mean squared error and R^2.

For the firth task ("Identify patterns/rules regarding delays and try to explain when delays should be expected, based on these patterns."), we used descriptive statistics to highlight that IAD airport presents the highest delays. Afterwards, we applied apriori algorithm to detect i) the most frequently delayed flight routes, ii) the most frequently delayed airlines based on the departure time and iii) the most frequently delayed airlines based on the day of the week. We extracted the respective rules in all cases.