NATIONAL CENTRE FOR SCIENTIFIC RESEARCH "DEMOKRITOS"

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΛΟΠΟΝΝΗΣΟΥ
UNIVERSITY OF PELOPONNESE

MSc in Data Science

2021-2022

Big Data Mining

Assignment 3

REPORT

Vitsas Alexandros-Konstantinos

Contents

1 Airlines dataset

1.1 Dataset overview

The dataset used in this task is in "arff" format. For the sake of convenience, the arff file was converted to a csv file with the use of an online file converter. The information obtained after the phase of EDA (Exploratory Data Analysis) are the following:

- The dataset consists of 7 attributes (columns) and 100161 instances (rows).
- There are 14766 duplicate rows.
- No missing values.

1.1.1 Attributes' description & interpretation

DayofWeek : categorical attribute with positive integer values ranging from 1 to 7. Each integer corresponds to a unique day of the week.

CRSDepTime : numerical attribute with positive integer values ranging from 5 to 2359. Each value corresponds to a departure time. So, for example, the value 1335 can be interpreted as $1:35pm$.

UniqueCarrier : nominal attribute with 9 distinct string values contained in the following list: ['UA', 'DL', 'CO', 'EA', 'NW', 'AA', 'US', 'TW', 'PA (1)']. Each value corresponds to a unique airline.

FlightNum : categorical attribute with positive integer values ranging from 12 to 4007. Each value corresponds to a unique flight. There are 624 distinct flight numbers in our dataset.

Origin : nominal attribute with 58 distinct string values contained in the following list: ['ORD', 'IAD', 'CLE', 'EWR', 'DTW', 'BOS', 'LGA', 'CAE', 'BDL', 'ISP', 'TPA',

’FLL’, ’RSW’, ’BUF’, ’ROC’, ’MCO’, ’JAX’, ’PBI’, ’GSO’, ’IAH’, ’DEN’, ’CHS’, ’SFO’, ’DFW’, ’LAX’, ’RDU’, ’PIT’, ’SLC’, ’JFK’, ’ATL’, ’PHF’, ’MSY’, ’MDW’, ’MIA’, ’MEM’, ’STL’, ’ORF’, ’SJU’, ’RIC’, ’DCA’, ’SEA’, ’MSP’, ’PHL’, ’MDT’, ’HPN’, ’BWI’, ’SRQ’, ’MHT’, ’SYR’, ’SDF’, ’BTV’, ’IND’, ’BNA’, ’HOU’, ’SAV’, ’PVD’, ’PWM’, ’PHX’]. Each value corresponds to a unique airport.

Dest : nominal attribute with 59 distinct string values contained in the following list:[’IAD’, ’ATL’, ’DEN’, ’RIC’, ’LGA’, ’MSP’, ’ORD’, ’DTW’, ’EWR’, ’SLC’, ’PIT’, ’BOS’, ’SJU’, ’CLE’, ’DFW’, ’LAX’, ’BDL’, ’FLL’, ’MCO’, ’SFO’, ’MIA’, ’MSY’, ’JFK’, ’PBI’, ’TPA’, ’CHS’, ’JAX’, ’PHF’, ’MEM’, ’MDW’, ’ROC’, ’BUF’, ’ISP’, ’GSO’, ’RDU’, ’ORF’, ’STL’, ’IAH’, ’CAE’, ’RSW’, ’SEA’, ’DCA’, ’PHL’, ’CRW’, ’HPN’, ’BWI’, ’SRQ’, ’MDT’, ’MHT’, ’SYR’, ’SDF’, ’IND’, ’BTV’, ’HOU’, ’SAV’, ’BNA’, ’PVD’, ’PWM’, ’PHX’]. Each value corresponds to a unique airport.

ArrDelays : numerical attribute with integer values ranging from -72 to 667. Positive values correspond to arrival delay. Negative or zero values could be interpreted as a flight arriving at its destination earlier than expected.

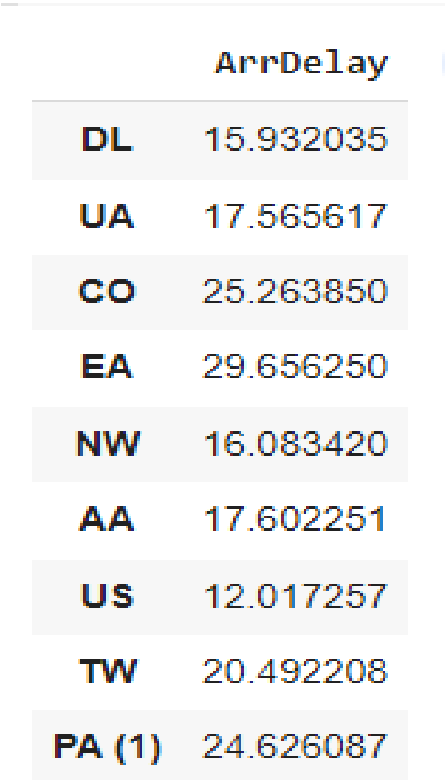
1.1.2 Observations & assumptions

After having a ’peek’ at the data, the following useful observations and subsequent assumptions are made:

- Duplicate instances are not considered as ’noise’ in our data and are, therefore, not dropped from the dataset for the rest of the tasks of this assignment.
- The value ’IAD’ exists in all of the instances either in the attribute *Origin* or in the attribute *Dest*. After a quick search, it was found that ’IAD’ represents the Washington airport. So, we can conclude that this dataset contains information for flights that depart or arrive at this airport. This observation will be taken into account later on.

1.2 Average delays

Being interested only in the delays of the flights in order to calculate the average delays per airline, we can consider only the rows of the dataset that have as a value a positive number in the *ArrDelay* column. So, we consider 48687 rows. The figure below illustrates the average delay (arithmetic mean) per airline. (Figure 1.1)



	ArrDelay
DL	15.932035
UA	17.565617
CO	25.263850
EA	29.656250
NW	16.083420
AA	17.602251
US	12.017257
TW	20.492208
PA (1)	24.626087

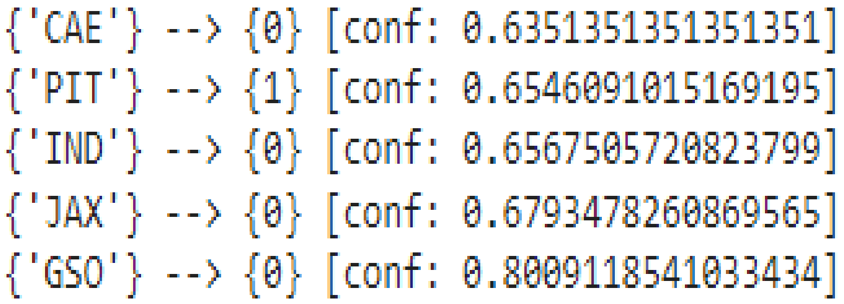
Figure 1.1: Average delay per airline in minutes Airline with maximum average delay: EA 29.7 minutes. Airline with minimum average delay: US 12.0 minutes.

1.3 Rules of association

We are interested in extracting rules of association between delays and point of origin and/or point of arrival. Therefore, only three features were considered for this task: *Origin, Dest, ArrDelay.* The feature *ArrDelay* was binarized, so that the value 0 corresponds to absence of arrival delay and the value 1 corresponds to presence of arrival delay.

Apriori algorithm was used for rule extraction. The observation mentioned in the previous section regarding Washington airport ('IAD') was taken into account in the process. More specifically, rules containing Washington airport either as a flight origin or as a flight destination are, more or less, useless. Also, whether an airport plays the role of the origin or of the destination of a flight, is not taken into account. So, for example, **"IF {'ATL'} THEN {1}"**, means that the airport 'ATL' is either the origin or the destination of the flight, which is likely to arrive with a delay.

After experimenting with the hyperparameters of the algorithm (support, confidence), some prominent rules are illustrated below. (Figure 1.2)



```
{'CAE'} --> {0} [conf: 0.6351351351351351]
{'PIT'} --> {1} [conf: 0.6546091015169195]
{'IND'} --> {0} [conf: 0.6567505720823799]
{'JAX'} --> {0} [conf: 0.6793478260869565]
{'GSO'} --> {0} [conf: 0.8009118541033434]
```

Figure 1.2: Rules of association between delays and point of origin and/or point of arrival with **minimum support** $= 0.005$, **minimum confidence** $= 0.6$

1.4 Delay prediction

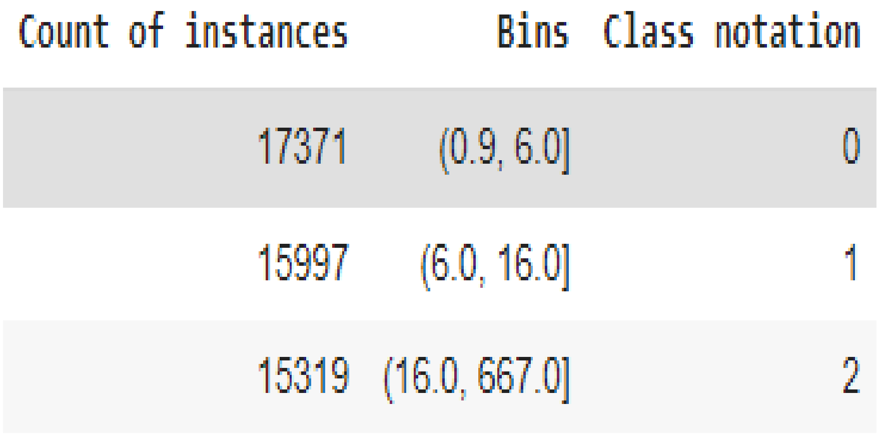
Delay prediction was handled as a classification problem, following the process described below:

- Only rows with positive delay values were considered.

- The attribute *FlightNum* was not considered for this task.

- Based on the observation regarding Washington airport ('IAD'), the attributes *Origin*, *Dest* were merged into one, by ignoring the existence of the value 'IAD' for each instance. So, four attributes are used to predict the delay: *DayofWeek, CRSDepTime, UniqueCarrier, Origin/Dest.*

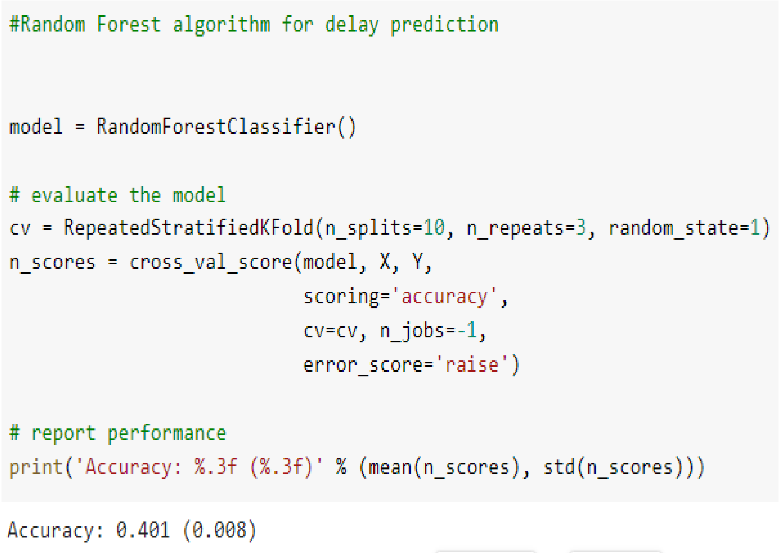
- The attribute *ArrDelay*, which plays the role of our target, was discretized into three states. Each state is represented by a bin, which corresponds to a different class in our classification problem. In order not to come up against any class imbalance problems, equal-frequency binning was applied. The results of the binning process are illustrated below. (Figure 1.3)



Count of instances	Bins	Class notation
17371	(0.9, 6.0]	0
15997	(6.0, 16.0]	1
15319	(16.0, 667.0]	2

Figure 1.3: Results of equal frequency binning of arrival delays.

- One-hot encoding was performed on categorical attributes which are not ordinal (*UniqueCarrier, Origin/Dest*). Days of the week, as well as departure times are assumed to have an ordinal nature and, thus, are not encoded. One-hot encoding produced 69 features which were used for the delay prediction.
- Random Forest classifier was used for the delay-class prediction. Its performance was evaluated with stratified k-fold cross-validation, as shown in Figure 1.4.



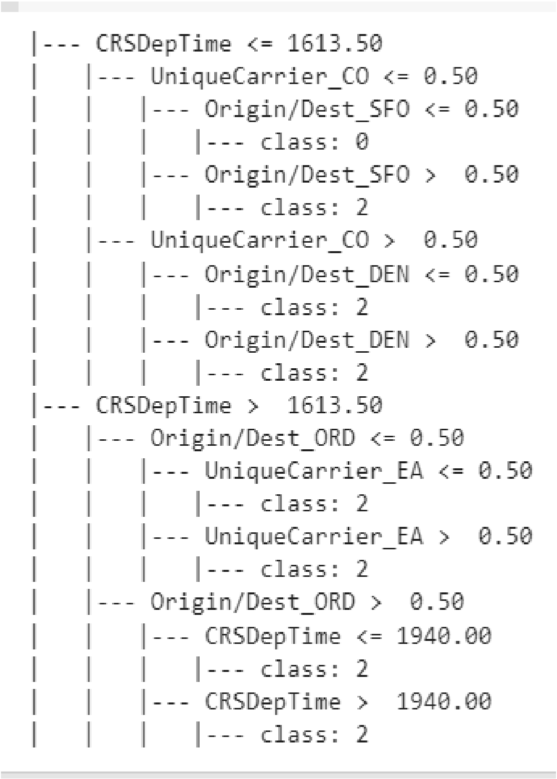
```
#Random Forest algorithm for delay prediction


model = RandomForestClassifier()

# evaluate the model
cv = RepeatedStratifiedKFold(n_splits=10, n_repeats=3, random_state=1)
n_scores = cross_val_score(model, X, Y,
                           scoring='accuracy',
                           cv=cv, n_jobs=-1,
                           error_score='raise')

# report performance
print('Accuracy: %.3f (%.3f)' % (mean(n_scores), std(n_scores)))
```

```
Accuracy: 0.401 (0.008)
```

Figure 1.4: Random Forest classifier performance on stratified k-fold cross-validation for $k = 10$

1.5 Rules regarding delays

A Decision Tree Classifier was used to extract rules regarding delays, utilizing the 69 features used in the previous task. Therefore, as non-delayed flights are not considered, rules are obtained regarding the length of the delay, according to the binning process described in the previous task. The figure below illustrates the results. (Figure 1.5)



```
|--- CRSDepTime <= 1613.50
|   |--- UniqueCarrier_CO <= 0.50
|   |   |--- Origin/Dest_SFO <= 0.50
|   |   |   |--- class: 0
|   |   |--- Origin/Dest_SFO >  0.50
|   |   |   |--- class: 2
|   |--- UniqueCarrier_CO >  0.50
|   |   |--- Origin/Dest_DEN <= 0.50
|   |   |   |--- class: 2
|   |   |--- Origin/Dest_DEN >  0.50
|   |   |   |--- class: 2
|--- CRSDepTime >  1613.50
|   |--- Origin/Dest_ORD <= 0.50
|   |   |--- UniqueCarrier_EA <= 0.50
|   |   |   |--- class: 2
|   |   |--- UniqueCarrier_EA >  0.50
|   |   |   |--- class: 2
|   |--- Origin/Dest_ORD >  0.50
|   |   |--- CRSDepTime <= 1940.00
|   |   |   |--- class: 2
|   |   |--- CRSDepTime >  1940.00
|   |   |   |--- class: 2
```

Figure 1.5: Rules extracted from a Decision Trees Classifier with maximum depth equal to 3.

For the given tree of maximum depth equal to (3), it seems that flights departing before 4 : 15pm, not belonging to the airline *CO* and not having the airport *SFO* as origin or destination are expected to have only a short delay (class 0). In any other case, the delay is expected to be long (class 2). The day of the week does not play a discriminative role for the given maximum depth of the tree.

2 Religion dataset

2.1 Data overview

The data used for this task were retrieved from an 'xls' file which was converted to a Dataframe. Fundamental information extracted from an initial examination of the dataset are the following:

- The dataset consists of 234 attributes (columns) and 3075 instances (rows).
- No duplicate rows.
- No missing values.

2.1.1 Attributes' description & interpretation

CNAME : nominal attribute with 3075 distinct string values. Each value represents the name of a unique US county.

STCODE : categorical attribute with positive integer values ranging from 1 to 56. Each value represents the *State Census Code* (demographic characteristic) of the respective US county.

TOTPOP : numerical attribute with positive integer values ranging from 52 to 4508792. Each value represents the total population of the respective US county.

TOTMEMB : numerical attribute with non-negative integer values ranging from 0 to 2685524. Each value represents the total number of religious members of the respective US county.

TOTCHUR : numerical attribute with non-negative integer values ranging from 0 to 1939. Each value represents the total number of churches of the respective US county.

XXX_M : attributes ending with ” _M” denote the number of religious members corresponding to religion ”XXX” for the respective US county. Such attributes are numerical with positive integers.

XXX_C : attributes ending with ” _C” denote the number of churches corresponding to religion ”XXX”. Such attributes are numerical with positive integers.

There are 114 religions present in the dataset.

2.1.2 Checking for inconsistencies

After examining the dataset for contradicting values the results are the following:

- There are 15 counties where the number of religious members is greater than the total population.
- There are no counties where the total number of members is not equal to the sum of religious members.
- There are no counties where the total number of churches is not equal to the sum of the churches of different religions.
- There are no duplicates either in terms of the counties or in terms of religions.

2.1.3 Visualizations

Some useful visualizations for summarizing the data are provided below:

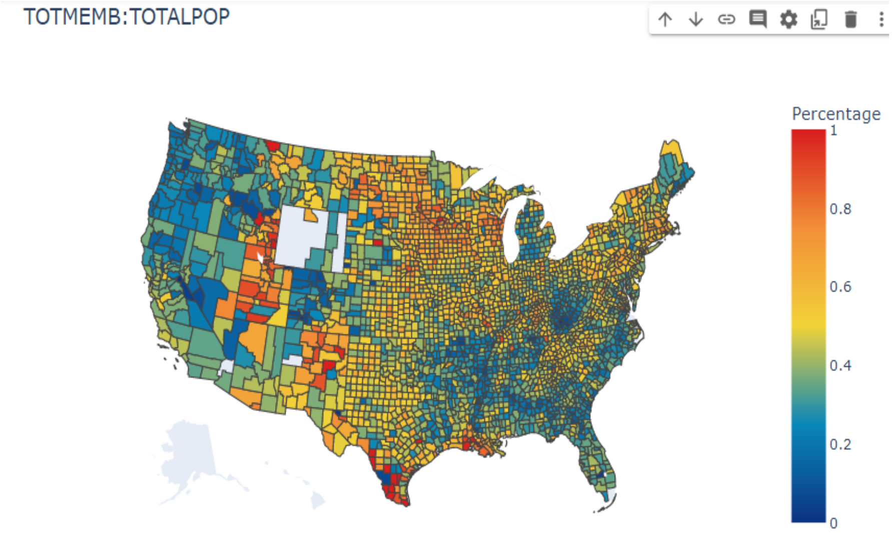


Figure 2.1: A map of the US depicting the percentage of the total population of each US county that is religious.

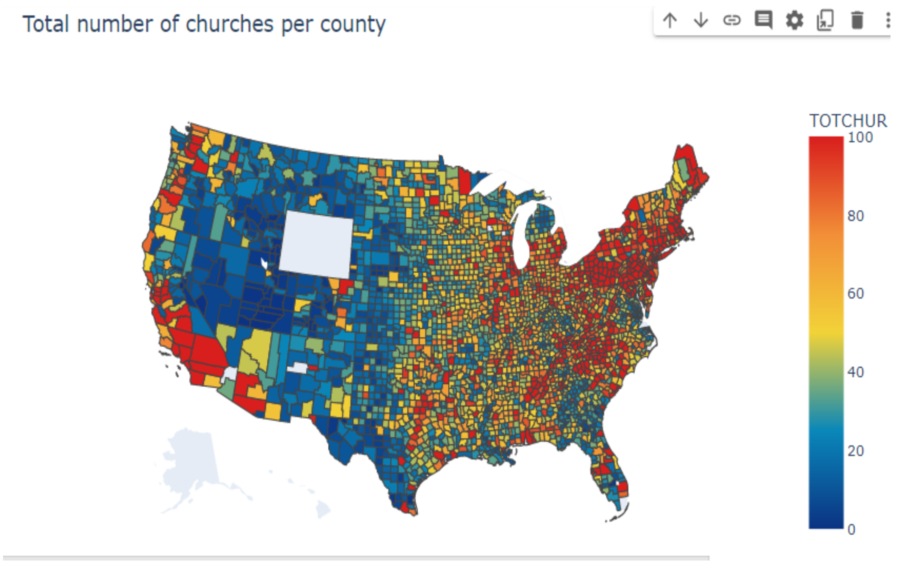
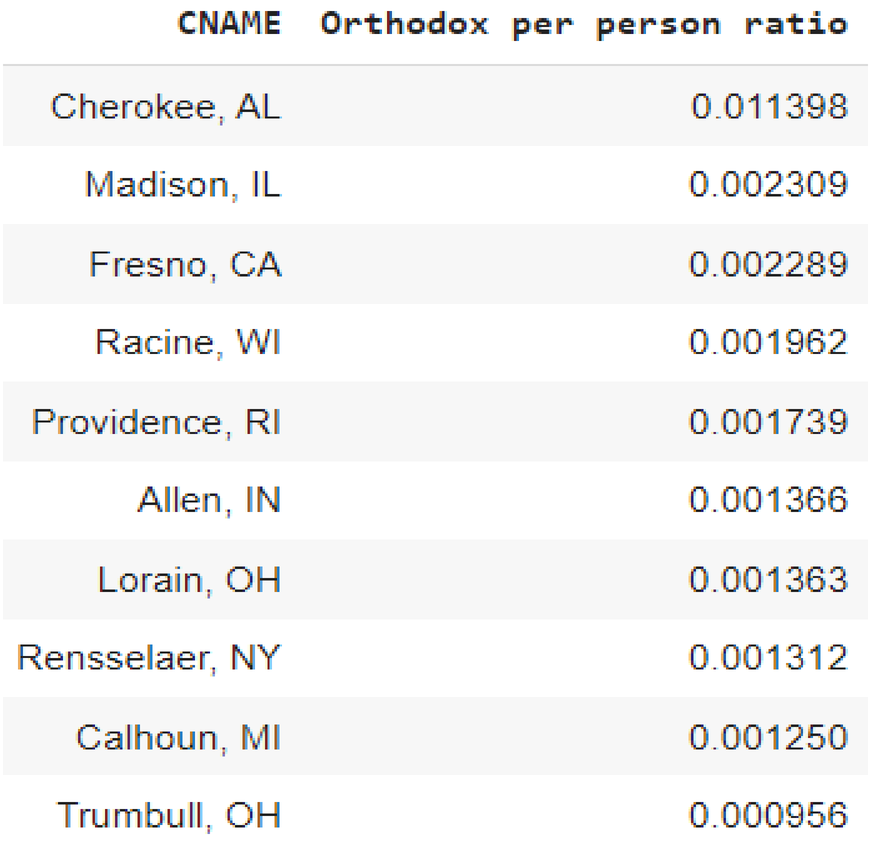


Figure 2.2: The total number of churches per county ranges from 0 to 1939. The mean value is approximately equal to 60 churches per county. In this map, counties having more than 100 churches are denoted with red color.

2.2 Orthodox Christian members

The data description provided in this link was used to find which of the religions present in the dataset refer to Orthodox Christians. For this task, it was assumed that ”Orthodox Christians” are simply mentioned as ”Orthodox”. The religions corresponding to Orthodox Christians, as denoted in the dataset, were found to be the following: *ARAPO, GRKAD, ACROC, BEOC*.

The per person ratio of Orthodox Christian members was calculated by dividing the total number of Orthodox members by the total population of each county. The top-ten counties with the highest per person ratio of Orthodox Christian members are presented below: (Figure 2.3)



CNAME	Orthodox per person ratio
Cherokee, AL	0.011398
Madison, IL	0.002309
Fresno, CA	0.002289
Racine, WI	0.001962
Providence, RI	0.001739
Allen, IN	0.001366
Lorain, OH	0.001363
Rensselaer, NY	0.001312
Calhoun, MI	0.001250
Trumbull, OH	0.000956

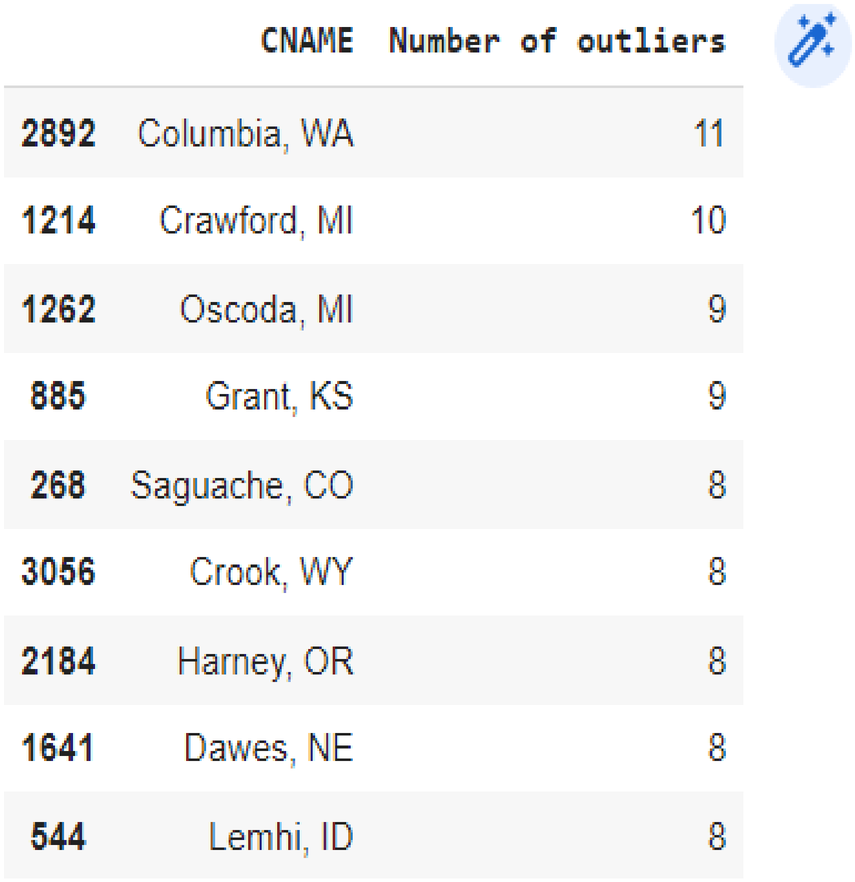
Figure 2.3: Counties with the highest per person ratio of Orthodox Christian members

2.3 Distribution of churches across religions - Extreme counties

The task of finding the 3 most extreme counties with respect to the distribution of their churches across religions was handled according to the following strategy:

- The distribution of churches across religions was considered for each county.
- Outliers of each of the above distributions were founded.
- The top-three counties with respect to outliers were considered as the most extreme counties with respect to the distribution of their churches across religions.

The results of this process are presented below. (Figure 2.4)



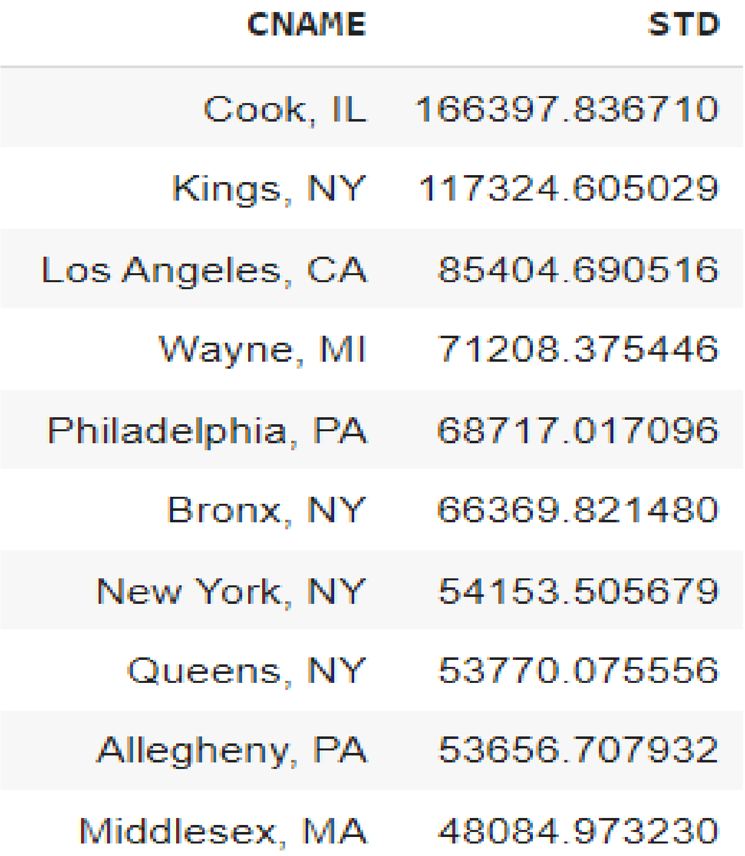
	CNAME	Number of outliers
2892	Columbia, WA	11
1214	Crawford, MI	10
1262	Oscoda, MI	9
885	Grant, KS	9
268	Saguache, CO	8
3056	Crook, WY	8
2184	Harney, OR	8
1641	Dawes, NE	8
544	Lemhi, ID	8

Figure 2.4: The top-ten most extreme counties with respect to the distribution of their churches across religions.

It seems that counties *Columbia, WA, Crawford, MI, Oscoda, MI, Grant, KS,* are the counties needed for the task.

2.4 Cross-religion center

In order to maximize the benefits from a cross-religion center, what should be taken into account is the pluralism with respect to religious members residing in a county. Following this rationale, the county with the greatest standard deviation with respect to the distribution of its religious members across religions is considered to be the ideal place to build a cross-religion center. Results of the analysis are illustrated below for the top-ten counties. (Figure 2.5)



CNAME	STD
Cook, IL	166397.836710
Kings, NY	117324.605029
Los Angeles, CA	85404.690516
Wayne, MI	71208.375446
Philadelphia, PA	68717.017096
Bronx, NY	66369.821480
New York, NY	54153.505679
Queens, NY	53770.075556
Allegheny, PA	53656.707932
Middlesex, MA	48084.973230

Figure 2.5: Top-ten counties with the highest standard deviation with respect to the distribution of their religious members across religions.

The results of this analysis indicate that county *Cook, IL* is the most appropriate county for this task.