Senior data scientist assessment

Data Pre-processing (clean/ETL):

1. Remove all the row records with null values of drop-off location
2. Select the data records in terms of actual GPS information of New York

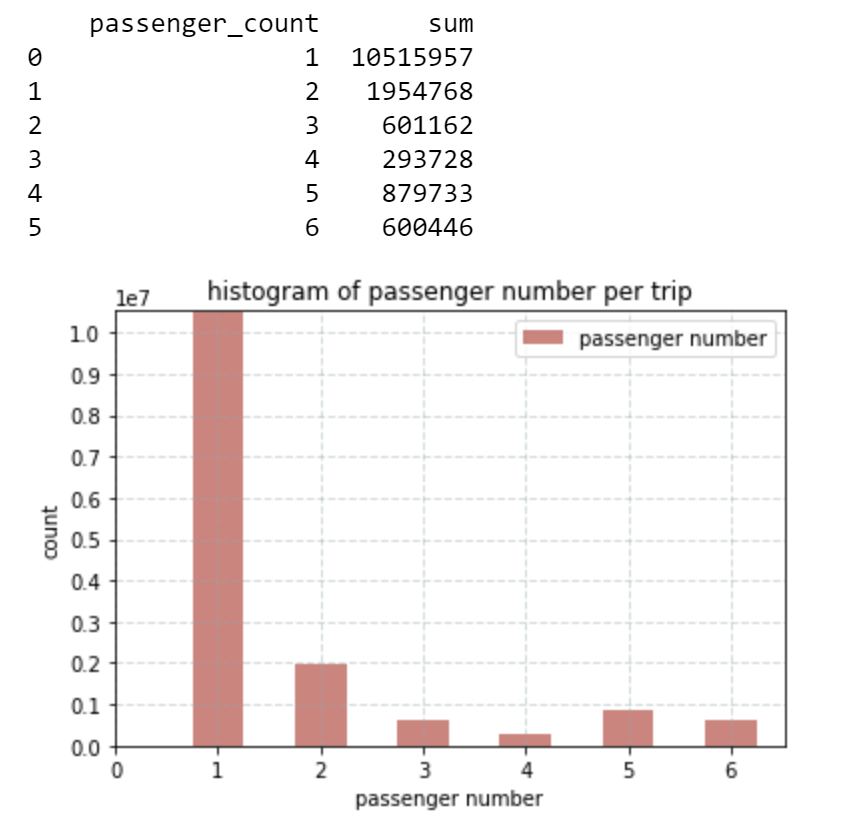
Data process platform:

* Ubuntu + python3.6 + Jupyter
* Scikit-learn, Pandas, numpy, matplotlib

Basic questions:

1. What is the distribution of number of passenger per trip?

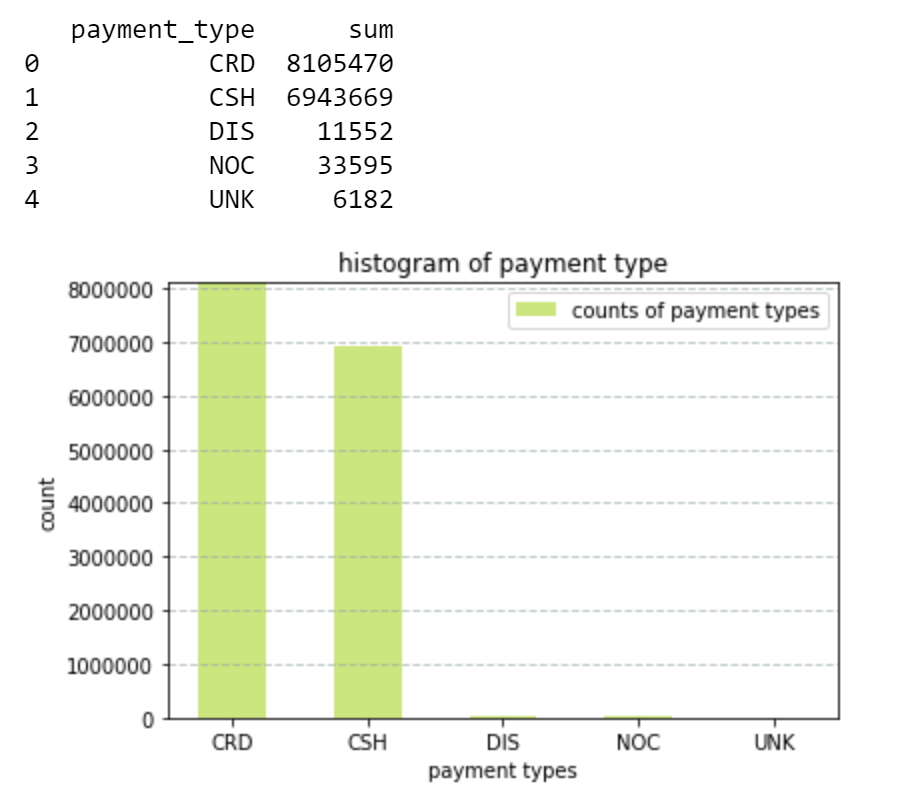
*In the raw data set, there are some records with 0 passengers or greater 7 passengers, which are outlier records. Therefore, I have removed them for analysing.*



From the result, we can see most of passengers take Taxi alone.

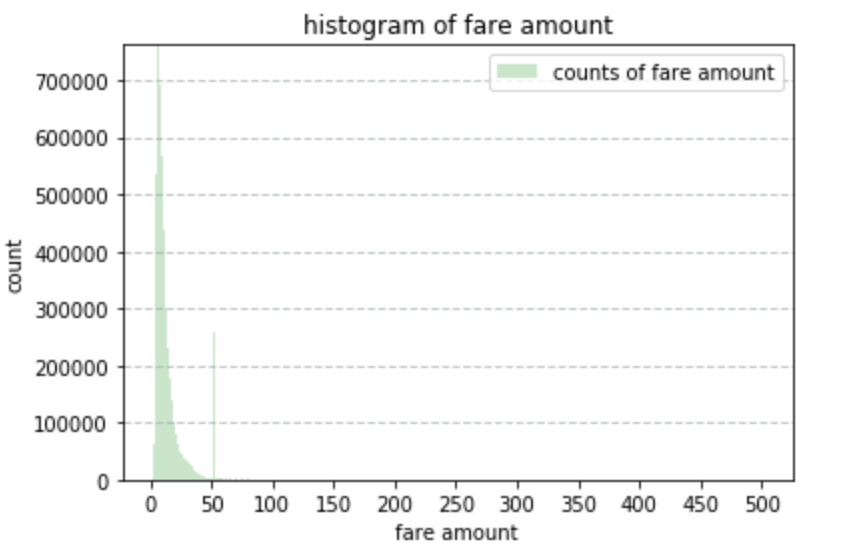
1. What is the distribution of payment type?

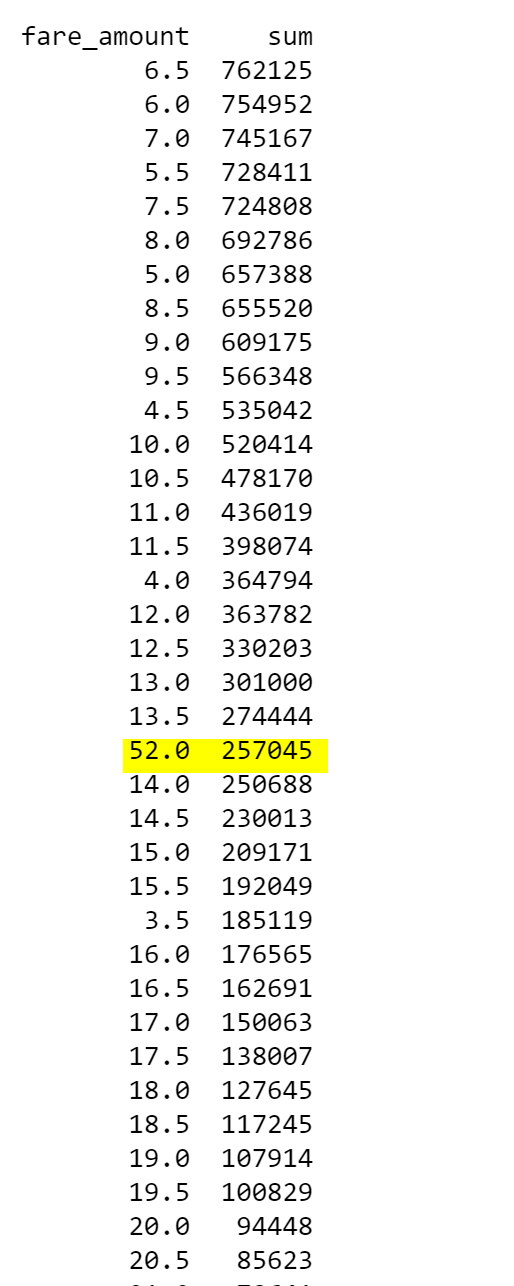
*More than half of passengers (53.68%) preferred to pay via credit card and 45.98% passengers paid by cash.*



1. What is the distribution of fare amount?

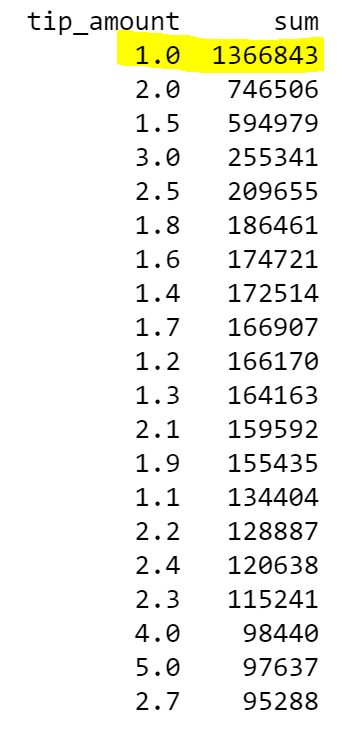
In terms of the frequencies of varying fare amounts in data set, most of passengers pay around 20 dollars especially from 4 dollars to 10 dollars.

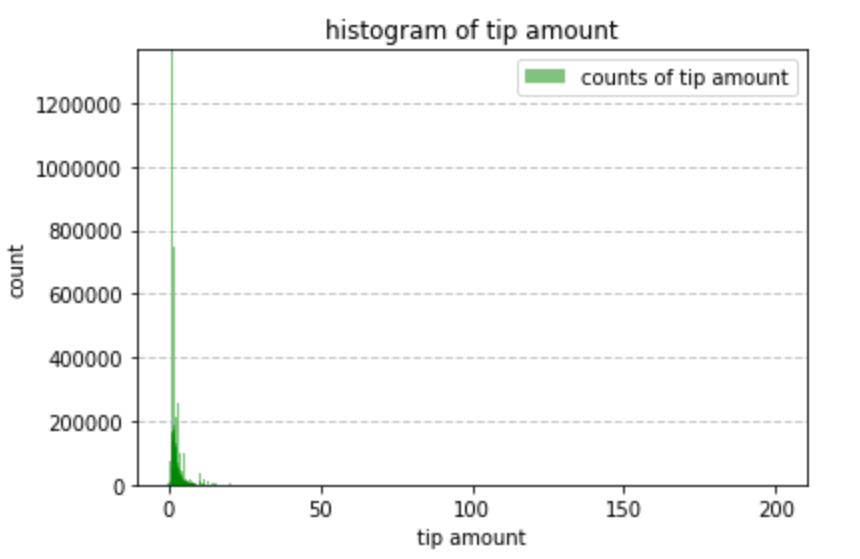




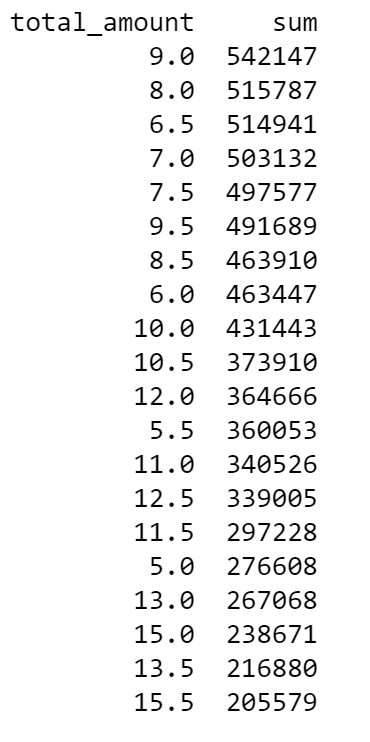
1. What is the distribution of trip amount?

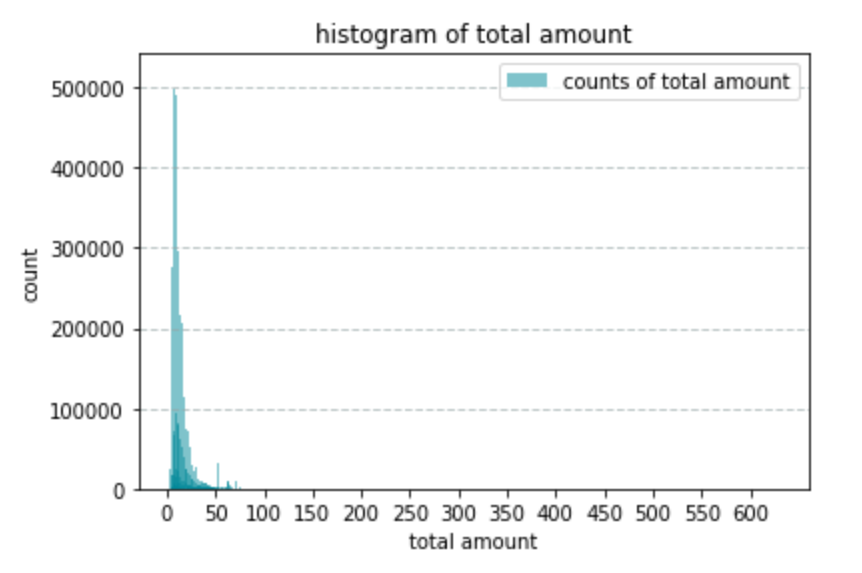
Passengers pay one dollar as tip most often. The range of tip amount is roughly between one dollar and five dollars.





1. What is the distribution of total amount?





1. What are top 5 busiest hours of the day?

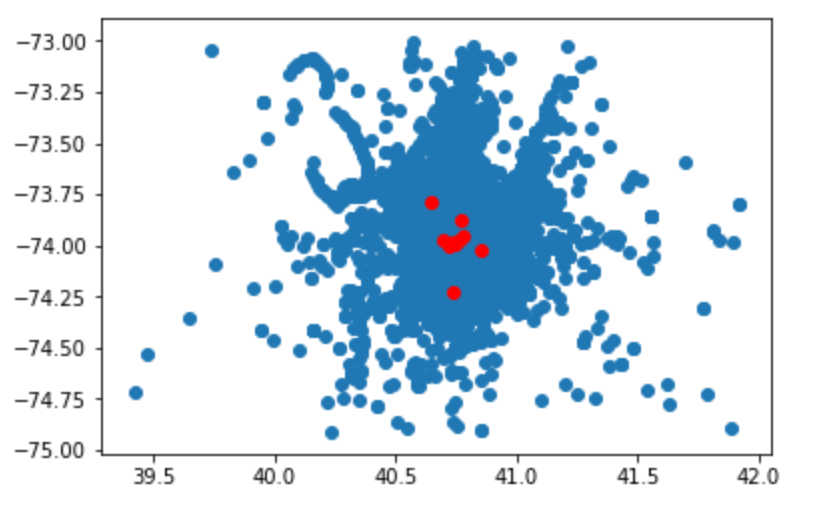
According to histogram of trips based on hours of the day, the top 5 busiest hours are:

|  |  |
| --- | --- |
| Time range | Number of trips |
| 19:00 - 20:00 | 950590 |
| 18:00 - 19:00 | 922177 |
| 20:00 - 21:00 | 917030 |
| 21:00 - 22:00 | 881281 |
| 22:00 - 23:00 | 852428 |

1. What are the top 10 busiest locations of the city?
2. Use kmeans to cluster the locations of pickup into 20 classes
3. Calculate the number of trips in each class
4. Sort the number and choose the 10 largest ones

|  |  |  |
| --- | --- | --- |
| latitude | longitude | Number of trips |
| 40.7618282 | -73.97609587 | 4315971 |
| 40.78291025 | -73.95836365 | 2737815 |
| 40.74939391 | -73.99531452 | 2579179 |
| 40.73560053 | -73.98586082 | 2299162 |
| 40.72097811 | -74.00346359 | 1901857 |
| 40.76843605 | -73.87352844 | 388169 |
| 40.69698248 | -73.97182165 | 366235 |
| 40.64676752 | -73.78528186 | 250143 |
| 40.85681094 | -74.0255909 | 2215 |
| 40.73834004 | -74.22646583 | 2022 |

The 10 busiest locations are marked as red spots as follow:



1. Which trip has the highest standard deviation of travel time?

I do not quite understand this question. “standard deviation” is related to a series of samples. If the column in trip\_time\_in\_sec is treated as a series of samples. The standard deviation is a number, so there is no “highest” standard deviation. However, if we can group the trips according to the trip distance and in each group we calculate the standard deviation, in this case, we can see when the trip distance is 57, the standard deviation is the largest.

1. Which trip has most consistent fares?

This question is like the above one. The result relies on the conceptions of “trip” and “consistent”. In general, when the standard deviation of a series of fares is small, we can say the corresponding “trips” are consistent. Otherwise, we can take use of variance to estimate it. First, calculate the mean of all fare amounts; second, check the fare amount in which trip is closed to the mean.

Open Questions:

1. In what trips can you confidently use respective means as measures of central tendency to estimate fare, time, taken, etc.
2. The trip record includes entire information, no null values
3. Use standard deviation to reselect the trips. For example, filter out the trips which
4. Can we build a model to predict fare and trip amount given pick up and drop off coordinates, time of data and week?
5. Join trip fare data and trip detail fare and get training data set <enrich\_data>
6. Build up training data. Input training data <train\_input> includes attributes pickup\_latitude, pickup\_longitude, dropoff\_latitude, dropoff\_longitude, time of day and week from <enrich\_data>, while output training data <train\_output> includes attributes fare\_amount and tip amount from <enrich\_data>
7. Train a Random Forest Regressor with train\_input and train\_output
8. Predict new test\_input and get prediction

Note: besides Random forest, MultipleLayer Perception or Neural Networks with a few hidden layers can also do it.

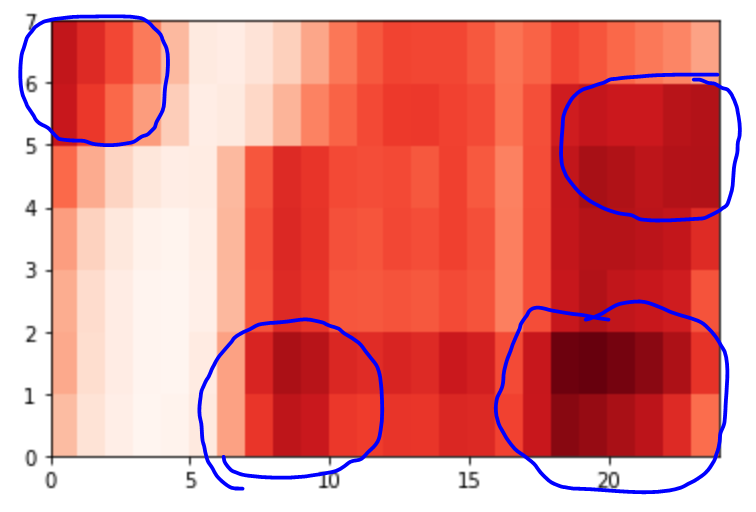
1. If you were a taxi owner, how would you maximize your earnings in a day?

The most important earnings for a driver is “tip amount”, so some factors must be considered to maximize earning in a day.

1. Rate of tip and time Tip\_Time => tip amount / trip time
2. Rate of tip and distance Tip\_distance => tip amount / trip distance
3. Relationship of tip and speed => calculate on which speed passenger is willing to pay more tip
4. On which location, the tip rate (tip / fare) is the most in terms of time

For a taxi owner, first find out the hot spots according to different time periods, and try to drive on the speed which passenger is willing to pay more tip,

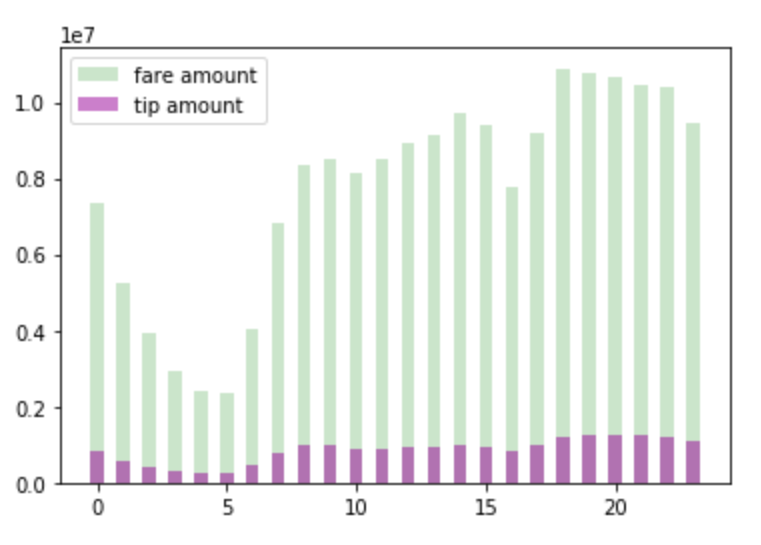
1. If you were a taxi owner, how would you minimize your work time while retaining the average wages earned by a typical taxi in the dataset?



Here is tip amount heatmap on hour of a day (x axis) and the day of week (y axis). For minimizing work time while retaining the average wage earned, a taxi driver should work during the hot spots (rush hours) which are shown in the heatmap.

1. If you run a taxi company with 10 taxis, how would you maximize your earnings?

Regarding 24 hours in a day, firstly we need to figure out “earning hours” in terms of fare amount and tip amount. For instance, set 0.75 x 1e7 as a threshold and define “earning hours” as “fare amount is greater this threshold”. In this case, all 10 Taxis should work in these “earning hours” as much as possible.



Next, we need to find out several hottest locations in the “earning hours”, “hottest” means more passengers are calling Taxi.

And then organize 10 Taxi drivers to cover the “hottest locations” as much as possible.

Show Case:

Publications:

* **Xiaoming Chen**, K. Fan, W. Liu, X. Zhang and M. Xue, *Discriminative Structure Discovery via Dimensionality Reduction for Facial Image Manifold*, Journal of Neural Computing and Applications 2015.

<https://link.springer.com/article/10.1007/s00521-014-1718-6>

* **Xiaoming Chen**, W. Liu, J. Lai, Z. Li and C. Lu, *Face Recognition via Locality Preserving Average Neighborhood Margin Maximization and Extreme Learning Machine*, Journal of Soft Computing 2012.

<https://link.springer.com/article/10.1007/s00500-012-0818-4>

* **Xiaoming Chen**, W. Liu, H. Qiu and J. Lai, *APSCAN: A Parameter Free Algorithm for Clustering*, Journal of Pattern Recognition Letters 2011.

<http://www.sciencedirect.com/science/article/pii/S0167865511000389>

* **Xiaoming Chen**, W. Liu, J. Lai and K. Fan, *Feature Extraction via Balanced Average Neighborhood Margin Maximization*, ICONIP 2011.
* <https://link.springer.com/chapter/10.1007%2F978-3-642-24958-7_13?LI=true>
* **Xiaoming Chen**, S. An, W. Liu and W. Li, *Object Detection by Admissible Region Search*, Advances in Artificial Intelligence. AI 2011. <https://link.springer.com/book/10.1007/978-3-642-25832-9#page=538>
* **Xiaoming Chen**, Z. Lu and Z. Li, *Video Retrieval Using VQ-based Global Motion Features*, VIE 2008.

<http://digital-library.theiet.org/content/conferences/10.1049/cp_20080380>

* X. Wu, **Xiaoming Chen**, L. Zhou, X. Li and J. Lai*, Adaptive Subspace Learning: An Iterative Approach for Document Clustering*, Journal of Neural Computing and Applications 2014.

<https://link.springer.com/article/10.1007/s00521-013-1486-8>

* J. Huang, **Xiaoming Chen**, PC. Yuen, J. Zhang, WS. Chen and JH. Lai, *Kernel Parameter Optimization of Kernel-based LDA Methods*, IJCNN 2008.

<http://ieeexplore.ieee.org/abstract/document/4634350/>

* Z. Li, KH. Yap and **Xiaoming Chen**, *Beyond Bag of Words: Combining Generative and Discriminative Models for Natural Scene Categorization*, ICASSP 2011.

<http://ieeexplore.ieee.org/abstract/document/5946566/>

* K. Fan, S. An, W. Liu and **Xiaoming Chen**, *Margin Preserving Projection for Image Set Based Face Recognition*, ICONIP 2011.

<https://link.springer.com/chapter/10.1007%2F978-3-642-24958-7_79?LI=true>

* S. An, P. Peursum, W. Liu, S. Venkatesh and **Xiaoming Chen**, *Exploiting Monge Structures in Optimum Subwindow Search*, CVPR 2010.

<http://ieeexplore.ieee.org/abstract/document/5540119/>

*Research area:*

* *Subspace learning for feature extraction (linear subspace learning, kernel based learning and manifold)*
* *Clustering*
* *Neural Networks*
* *Computer vision and pattern recognition*