**Trip Destination Prediction Problem**

**Introduction**

We want to solve the Trip Destination Prediction problem [1]. This problem needs to predict the trip destination in a new area, called Kinki in Japan by using the four other Japanese zone trip trace records. This proposal includes 4 parts. We will list our plan step by step in “Our Method” and explain how we get it in “Our plan”.

**Reason we choose it**

We are highly interested in the topic. By using a simple assumption in one paper you can get an accuracy of 0.43 [2]. We believe there is much space left to be improved in this problem. Another reason is this topic is very suitable for machine learning and data analysis research. We can better study and learn some data mining content through this research.

From a social perspective, this topic can better predict and improve the efficiency of public transportation systems and help us understand human travel patterns. This is of great significance to transportation planning and social development and can also provide more powerful data support for decision makers.

**Our plan**

After examining the data closely, we can separate this problem into two sub-prediction problems. The first is classifying whether the traveller will make a round trip. The second is predicting the destination if the traveler does not make a round trip.

The concept of a “round trip” is put forward by *“An Embarrassingly Simple Rule-based Visiting Circulation Approach to Trip Destination Prediction*” [2]. The round trip in this context is based on a simple assumption that most people will return to their origin after they finish their trips. In this paper, they sort the test data according to their “Pid” and “Departure\_time”. If there is only one record in the same day, then just set “Destination” as the same as their “Origin”. If there are more than one record on the same day, set the destination of the next one as the origin of the previous one, and set the origin of the first one as the last one. The assumption is intuitive, as most people depart from and return to their home at the end of a trip. With just a very simple method, they achieved an accuracy of 43% [2]. It turns out that their assumption is correct, and the dataset is completed. It includes the full trip trace of both the beginning and end. However, 57% of the records are still not round trip.

We will improve this by introducing two models. The first model is a round-trip classification model, using personal information such as age, trip type, etc., to classify whether the traveler would make a round trip. The second model is the destination prediction model, which uses geographical information to predict the destination of a non-round trip.

For the second model, We can combine with the method from “*A Hybrid Method with Gravity Model and Nearest-Neighbor Search for Trip Destination Prediction in New Metropolitan Areas*” [3] using the geographical information like gravity model [3], and using the origin in the test data [3] to identify traffic junctions and metropolitan areas in Kinki. We can also use 25 kilometres [3] as our middle destination radius, and develop a model to score candidate zones to select the optimal one.

**Our Method:**

1. Clean the train data by dropping the row where the label does not appear in the test data. For example, like Trip Type “To Travel”, and adapt the “Occupation” label in the training data to match the test data.
2. Get train data for the round-trip classification model. Using the data which its origin of record does match its final destination.
3. Train the round-trip classification model using the data from step two.
4. Get train data for the round-trip classification model.
   1. Using the “Origin” to estimate metropolitan areas and other metrics, described in the paper “A Hybrid Method” [3].
   2. Using the trip type, age, and other information in the trip records.
   3. Using data in the shapefile and zone feature files.
5. Using the data in step four to train the destination prediction model.
6. Fill in the Kinki destination with “Origin” for the round trip given by model one and result of the model two for the non-trip.

# **Exploratory data visualization**

The dataset includes two parts of data. The first part is the travel records. Those data contain information about the traveler and the trip proposal. This part of data includes:

* Pid
* Departure\_time
* Gender
* Age
* Occupation
* Trip\_type
* Origin

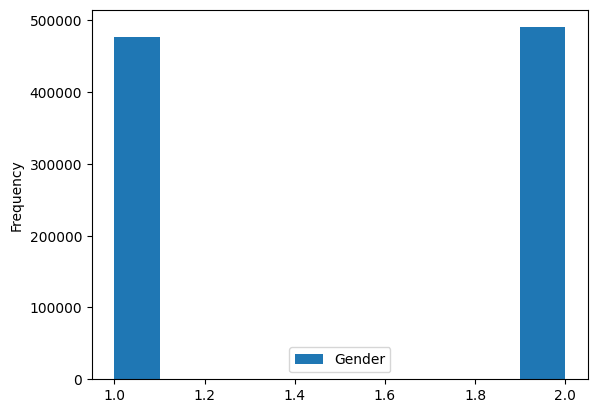
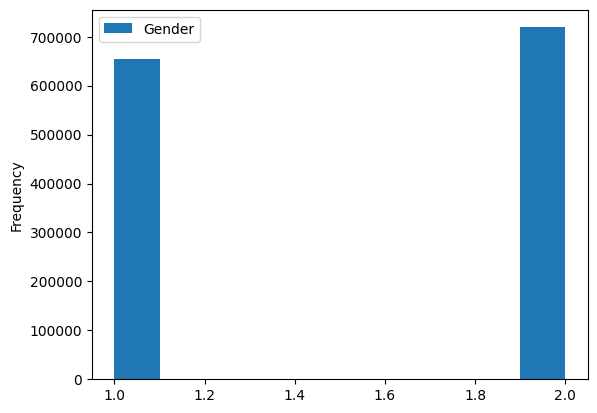
The train data has the label **Destination**, which is our target.

The train data is from four areas in Japan. They are Tokyo which contains 790,613 trip records, 470,140 travelers and 1655 zones; Chukyo which contains 205,435 trip records, 135,910 travelers and 1009 zones; Kyushu which contains 343,747 trip records, 147,968

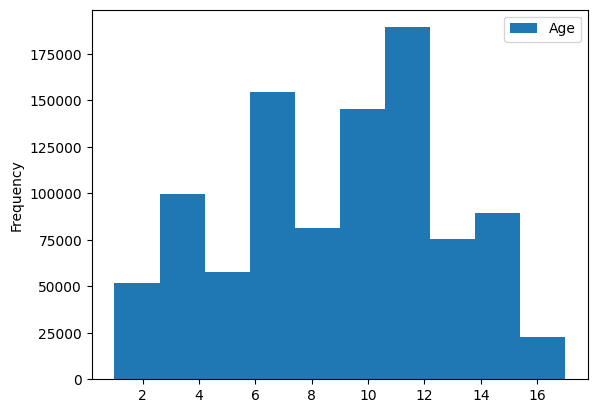
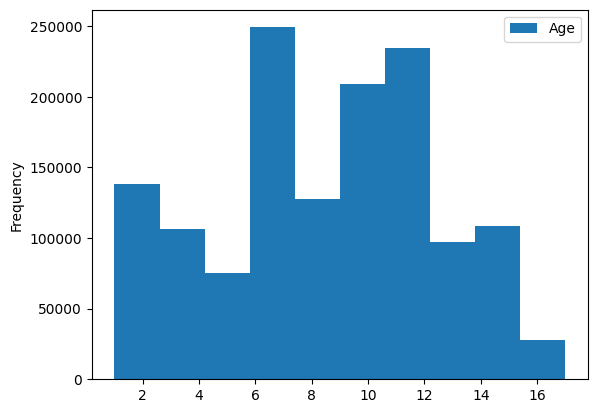
travelers and 565 zones; Higashisuruga which contains 34,496 trip records, 21,968 travelers and 78 zones. And our test data Kinki contains 967,459 trip records, 624,511 travelers and 1080 zones. Because of the specificity of geography, it is better to choose a whole area as our validate dataset. Considering the data scale, Tokyo is the most similar area to our test data. **So, we will use Tokyo to validate our model.**

First, let’s check the distribution of train data between the test data.

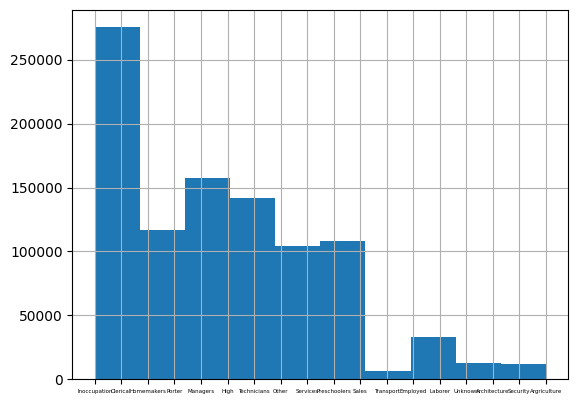
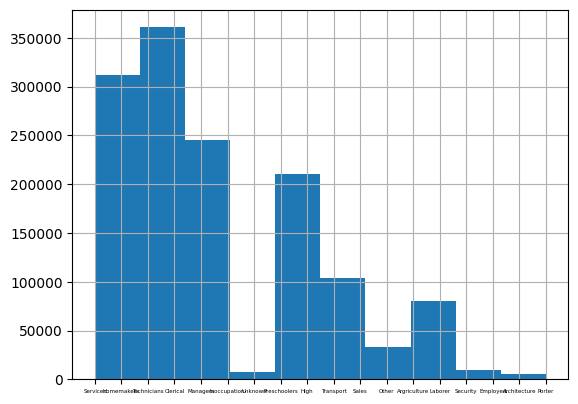
**Gender-train data Gender-test data**

****

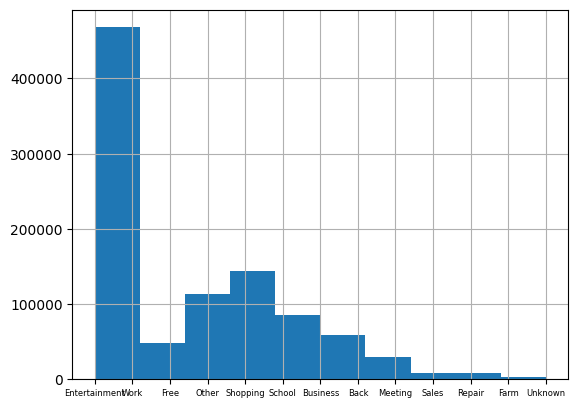
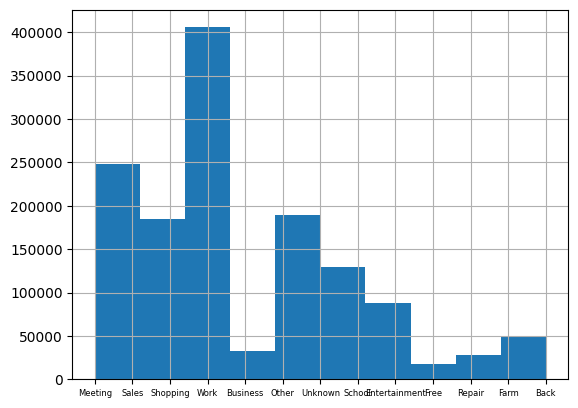
**Age-train data Age-test data**

****

**Occupation-train data Occupation-test data**

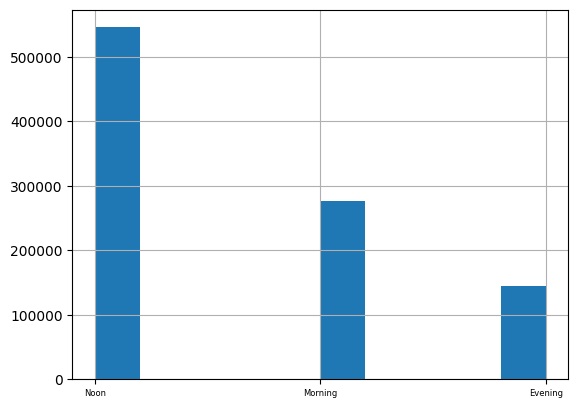
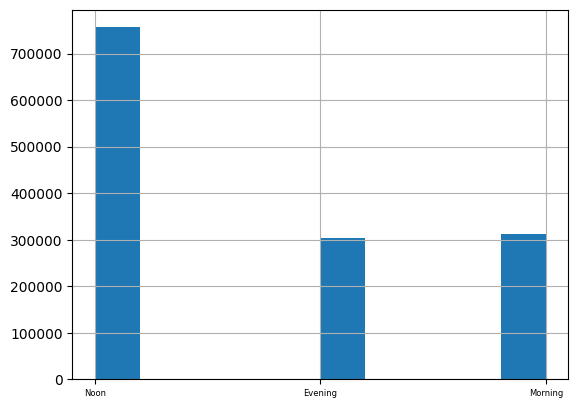
****

**Trip Type-train data Trip Type-test data**

****

We also map the departure time into 3 labels:Morning(0:00~7:59);Noon(8:00~15:59); Evening(16:00~23:59).

**Departure Time- train data Departure Time- test data**



We can learn from the graph that both the Trip Type and the Occupation have a significant difference between the train data and the test data. The most trip type in the train data is “To work” (29%) while the most frequent trip type in test data is “To Eat/Social/Entertainment”(13%). We need to reinforce the “To Eat/Social/Entertainment” labels in our training and reduce the “To work” labels.

We also look at the label distribution between round-trip and non-round-trip. We found that Trip Type and Occupation also have significant differences between round-trip records and non-round-trip records.

|  |  |  |
| --- | --- | --- |
| **Trip Type Label** | **round(%)** | **non-round(%)** |
| Farm Chores | 93.05273 | 6.947266 |
| Back to Work/School | 90.14294 | 9.85706 |
| For Sales/Delivery/Restock/Purchase | 88.00734 | 11.99266 |
| For Meeting/Withdraw/Doctors Visit | 85.93709 | 14.06291 |
| Repair | 81.95917 | 18.04083 |
| Other Business Purpose | 70.25719 | 29.74281 |
| Unknown | 65.0745 | 34.9255 |
| Other Private Purpose (including To Hospital and Pickup) | 57.66347 | 42.33653 |
| To Eat/Social/Entertainment | 53.97089 | 46.02911 |
| To Shopping | 53.15415 | 46.84585 |
| To School | 50.52002 | 49.47998 |
| Free Purpose | 47.47218 | 52.52782 |
| To Work | 45.01769 | 54.98231 |
| To Home | 0 | 0 |

|  |  |  |
| --- | --- | --- |
| **Occupation label** | **round(%)** | **non(%)** |
| Argriculture Forestry and Fishery | 72.87994 | 27.12006 |
| Sales | 67.73894 | 32.26106 |
| Homemakers | 67.06349 | 32.93651 |
| Inoccupation | 66.55221 | 33.44779 |
| Preschoolers Elementary and Middle School Students | 64.98938 | 35.01062 |
| Other Occupation | 62.36106 | 37.63894 |
| Transport | 61.11602 | 38.88398 |
| Services | 59.22392 | 40.77608 |
| Laborer | 59.03273 | 40.96727 |
| Managers | 58.25114 | 41.74886 |
| Security Inspectors | 57.09009 | 42.90991 |
| Clerical Assistants | 55.27876 | 44.72124 |
| Technicians and Professionals | 55.04421 | 44.95579 |
| Employed(Detail Unknown) | 49.10714 | 50.89286 |
| Unknown | 48.7385 | 51.2615 |
| Architecture and Mining Workers | 46.07081 | 53.92919 |
| Porter Cleaner and Packer | 38.18304 | 61.81696 |
| High School Students and above | 26.02425 | 73.97575 |

From the label “**trip type**”, we can see most of “Farm Chores”,”Back to Work/School”,”For Sales/Delivery/Restock/Purchase”,”For Meeting/Withdraw/Doctors Visit”,”Repair” are **round trips.**

From **“Occupation”** label, it is likely that the people who engage in “Argriculture Forestry and Fishery” would take a round-trip, while the people who from “High School Students and above” would take a non-round-trip.

The second part is the geographical data. It can also be divided into two parts. One is the human geography part. It contains:

* ZONE\_ID
* Number of business establishments
* Number of employees
* Number of business establishments
* Number of employees
* Night Population

Because most of the human geographical data is continuous, we can calculate the pairwise correlation of those data.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| index | ZONE\_ID | T000918002 | T000918006 | T000918021 | T000918025 | T000847001 |
| ZONE\_ID | 1 | -0.16269 | -0.08568 | -0.12918 | -0.09552 | -0.18814 |
| T000918002 | -0.16269 | 1 | 0.653958 | 0.705365 | 0.568433 | 0.594927 |
| T000918006 | -0.08568 | 0.653958 | 1 | 0.713272 | 0.912627 | 0.367312 |
| T000918021 | -0.12918 | 0.705365 | 0.713272 | 1 | 0.809929 | 0.280525 |
| T000918025 | -0.09552 | 0.568433 | 0.912627 | 0.809929 | 1 | 0.215985 |
| T000847001 | -0.18814 | 0.594927 | 0.367312 | 0.280525 | 0.215985 | 1 |

There is a strong correlation between T000918006(Number of employees (secondary sector of industry)) and T000918025(Number of employees (tertiary sector of industry)). Maybe we could add them or only choose one of them in our further classification.

The other is geospatial data. In it are the following:

* ZONE\_ID
* POLYGON

POLYGON is a class which contains the information of area center and border. We can use this class to draw a plot to visualize our human geographical data and other data related to geography.

Let’s first look at the frequency between the origin and the destination.

|  |  |  |
| --- | --- | --- |
| **Tokyo** | **origin\_freq** | **dest\_freq** |
| **origin\_freq** | 1.000000 | 0.813921 |
| **dest\_freq** | 0.813921 | 1.000000 |

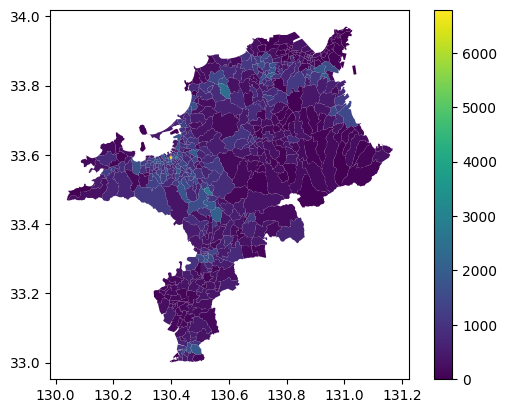
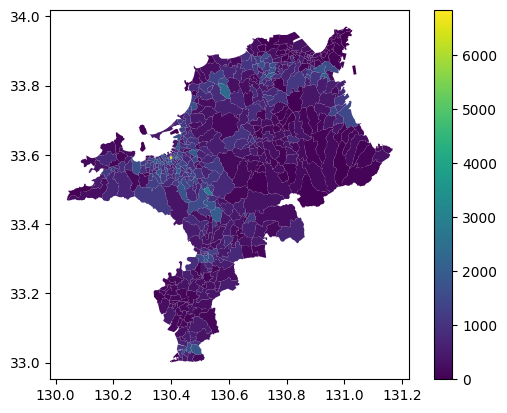
|  |  |  |
| --- | --- | --- |
| **Chukyo** | **origin\_freq** | **dest\_freq** |
| **origin\_freq** | 1.000000 | 0.743918 |
| **dest\_freq** | 0.743918 | 1.000000 |

|  |  |  |
| --- | --- | --- |
| **Kyushu** | **origin\_freq** | **dest\_freq** |
| **origin\_freq** | 1.000000 | 0.996793 |
| **dest\_freq** | 0.996793 | 1.000000 |

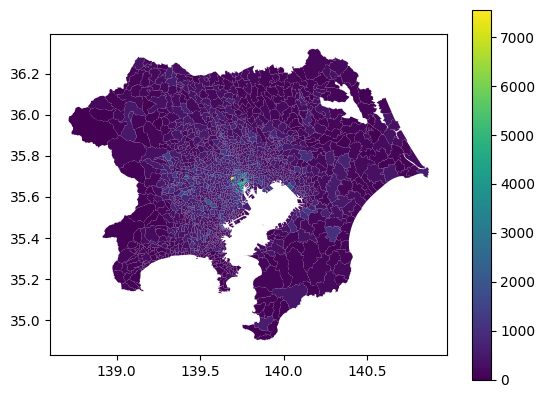
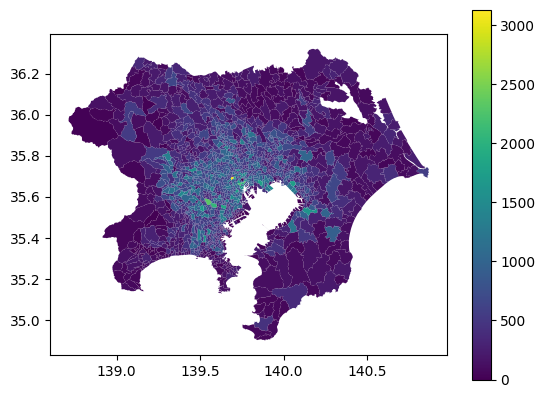
|  |  |  |
| --- | --- | --- |
| **Higashishuruga** | **origin\_freq** | **dest\_freq** |
| **origin\_freq** | 1.000000 | 0.908387 |
| **dest\_freq** | 0.908387 | 1.000000 |

The popular origin is also the popular destination. We can also use the origin in the test data as a candidate for our predication. We can visualize the frequency of Tokyo and Kyushu as following:

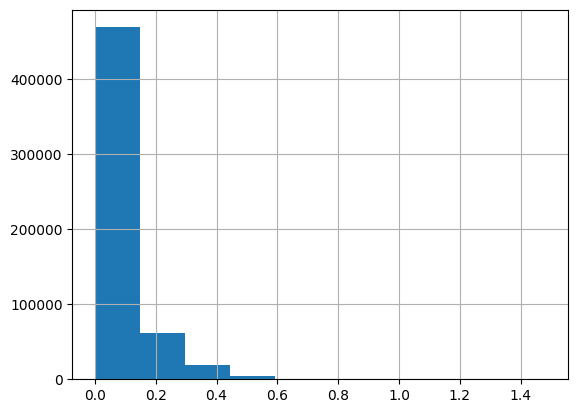
**Freq of Origin in Kyushu Freq of Destination in Kyushu**



**Freq of Origin in Tokyo Freq of Destination in Tokyo**



According to train data, over 77.86% of destinations are within 0.1 radius of the origin.



We can use 0.1 as our radius to choose the candidate zones for our predication model.

# **Round-Trip classification**

Because there is a lot of Categorical data and it is a classification problem, I would like to try Decision Tree and Logistic Regression.

From the previous exploratory, we will remove the "T000918025" and only train the data whose Trip\_type is unequal to 9. Trip\_type 9 would be set to round-trip trips.

Here is our model performance:

**Accuracy**

|  |  |  |
| --- | --- | --- |
|  | Decision Tree | Logistic Regression |
| train accuracy | 88.29% | 55.83% |
| validate accuracy(Tokyo) | 83.00% | 51.40% |

Because it is easy to predict the destination for a round-trip, when accuracy is at the same level we hope the [**False positive**](https://en.wikipedia.org/wiki/False_positive) to be minimum.

|  |  |
| --- | --- |
| **Confusion matrix**  Decision Tree on validate data | |
| 341956 | 80265 |
| 54102(FP) | 314290 |

|  |  |
| --- | --- |
| **Confusion matrix**  Logistic Regression on validate data | |
| 62956 | 51092 |
| 333102(FP) | 343463 |

Decision Tree is much better than Logistic Regression no matter on accuracy or False positive on this dataset. We would choose **Decision Tree** as our classification model.

# **Destination Prediction**

To be continued.

**Reference**

[1] *E.-S. Tu, Y.-H. Chen, E.-C. Liu, H.-Y. Keng, and C.-T. Li, “An Embarrassingly Simple Rule-based Visiting Circulation Approach to Trip Destination Prediction,” in 2022 IEEE International Conference on Big Data (Big Data), Dec. 2022, pp. 6565–6572. doi:* [*10.1109/BigData55660.2022.10020650*](https://doi.org/10.1109/BigData55660.2022.10020650)*.*

[2] *M.-H. Li, B.-Y. Chen, and C.-T. Li, “A Hybird Method with Gravity Model and Nearest-Neighbor Search for Trip Destination Prediction in New Metropolitan Areas,” in 2022 IEEE International Conference on Big Data (Big Data), Dec. 2022, pp. 6553–6560. doi:* [*10.1109/BigData55660.2022.10020439*](https://doi.org/10.1109/BigData55660.2022.10020439)*.*