**The Himalayas prediction Project**

# **Group 4 Members: Jingwen Cai, Shixue Mai, Yuping Liao, Zhenyu Cheng**

# **Statement of the Research Problem**

# The Himalayas are famous for their complex geography and the world's highest peak, Mount Everest (8848.86m). With 468 peaks flagged in the Himalayan database, the challenge and variety of climbing make it a mecca for extreme challengers. It attracts nearly 700,000 visitors every year, including experienced professional climbers, but unfortunately, the extreme altitude and bad weather took countless lives. The Annapurna Massif has been named the world's most dangerous mountain with its 29.5% fatality rate.

# In recent years, the population profile for climbing is dynamic. More and more people join the team no matter the race, nationality, age and gender. Tour companies can adjust their target customer accordingly by offering different climbing packages. Insurance companies can add the population profile into their consideration as well, by provide climbers different services and features.

# We wanted to use the Himalayan database compiled by Elizabeth Hawley to explore the factors necessary for a successful summit, the core causes of death, the target customer groups, and build a predictive model for the outcome of the journey. With this model, we are able to advise amateur climbers and tour companies on budget allocations, altitude predictions, and pre-departure preparation advice. At the same time, our predictive model can also serve insurance companies, helping them to estimate costs more accurately, provide rescue assistance, and allocate resources to maximize the economic benefits.

# **Literature Review**

# The Himalayan database consists of the expedition records for all climbed in the Nepal Himalaya. Mount Everest is an extreme environment, and sadly death on Everest occurs almost every year. The cause of death is diverse for all climbers, such as exhaustion, fall, disappearance, etc (Salisbury, 2021). The primary cause of death on Everest was Avalanche and Fall between 1950 and 2019. Death rates by peak altitude and season have also been well researched in this article, that the altitude range from 8000 to 8499m has the greatest death rate, while the majority of death happened during winter.

# In addition to death analysis, other literature suggests that the death rate changes from time to time (Huey, 2020). In this study, the researchers conducted the recent pattern (2006-2019) of success rate and death rate on Everest and compared it with 1990-2005. They found that the probability of summiting has increased dramatically, and more and more women joined the team. The climbers' backgrounds are dynamic, and we want to provide an analysis of the current environment. These two studies gave us some background information that support our research questions, that is how we can explore the safest team composition and achieve success.

# Different from the prior research, our research is focusing on the variables that are correlated with team composition, such as the proportion of leaders, the proportion of Sherpa/Tibetan and the proportion of speed climbed members, etc.

# **Research Question Outline**

# 1. What are the key factors for a successful summit

# Based on the literature review, which gave us a pitch idea of factors that can lead to successful ascents, and which factors neglected can lead to serious failure or even death, such as the composition of team members, tools used, weather, season, etc.

# In this project, we will use historical data to assess the importance of these attributes, the relationship between factors, and their combined impact on outcomes. After an evaluation, we may be able to answer the question of whether Sherpas are necessary, and whether a solo attempt is impossible.

# 2. What is the highpoint they can achieve and if they can succeed given their conditions

# After finding the key factors, we want to build a predictive model that can predict the outcome of the expedition. By entering pre-departure data, we can get the possible journey results such as if they can succeed and how high they could reach the mountain, we could give a pre-departure feedback to tell if the customer is suggested to climb based on the historical data and if the highpoint predicted within their expectations for further adjust on the team combinations.

# **Data Description**

# The original datasets are retrieved from the Himalayan database, and there are totally four tables, which are expeditions, peaks, references, and members. The ‘expeditions’ table contains information about all the expeditions that have climbed in Nepal or on its border peaks. The ‘peaks’ table displays all the information about all the mountaineering peaks of Nepal. And the ‘members’ table records all the members in each expedition. The ‘references’ table is added for literature references for those expeditions. The original data structure is shown in fig1. For this project,we only used the data from expeditions and members tables as they cover all the information, we need to predict whether one expedition will succeed or not. Moreover, due to the development of science and technology, the invention of many new instruments has greatly increased the success rate of summiting. In order to better build a prediction model, we decided to only use the data from 1950-2015 to control this variable. After preprocessing and combining the two datasets, we have a dataset containing 9009 observations and 44 variables where each observation represents one expedition in history. The variable description is shown in the following table.

The variables labeled in blue can be considered as dependent variables. From these variables, we can have a more comprehensive picture of one expedition regarding multiple aspects. For instance, did they succeed? How long did it take for them to reach the highest point? How many of them survived? Among those who survived, how many of them are the staff? These are all very intriguing problems, and we want to investigate this dataset from as many perspectives as possible.

**Data Cleaning Summary**

As mentioned before, each observation from the original ‘expeditions’ table records each operation, however, each observation from the original ‘members’ table is a separate member. To combine the two tables, we aggregate the data in the ‘members’ table based on the expedition id and join the result to the ‘expeditions’ table on expedition id. Specific code would be attached to the submission in the form of an R markdown document.

Moreover, we conducted data binning on sponsor, nation, leader, and route variables based on their occurrence. As higher occurrence indicates more experience in participating expeditions, we believe binning on these variables are necessary.

**Analytical Technologies**

As mentioned above, to figure out the most important factors we need to consider, we have done lasso regression analysis on both if the climber can succeed and how high they could reach and returned top variables on our dependent variables. We will determine those variables as a must for our customer to fill in so that we can have a better prediction. To build our model, for the classification part, we predict if the climber can succeed using logistic regression and gradient boost machine in classification Bernoulli model as well as neural networks implemented in gradient boost machine. For the general regression prediction, we use the linear regression to get a first intuition of the model and we transfer to a gradient boost machine in gaussian distribution and apply them into the h2o package for more accurate prediction. We tuned our model with different parameters so that we could have a lower RMSE for highpoint prediction and higher accuracy for classification.

**Analysis Results**

**Tourist Companies**

From the trend analysis, we find that Japan, USA, UK, France, and Spain have larger customer base than other countries, as shown in fig2. The top five of most popular peaks are Everest (EVER), Cho Oyu (CHOY), Ama Dablam (AMAD), Manaslu (MANA), Dhaulagiri I (DHA1), as shown in fig3. We also plot the age distribution of customers from the mentioned above five countries against the time, as shown in fig4. The plot indicates that as time goes by, climbers' age range expands significantly, as well as elder participate into the activity. The plot suggests that companies that provide services for climbers can shift their target customer on older people. Furthermore, we graph the age distribution of the top 5 popular peaks and structural differences are detected, as shown in fig5. Majority climbers of Dhaulagiri I are relatively young compared to others, suggesting that tourist companies should conduct specific research and adjust their strategy accordingly. We conducted a time-series analysis of the number of expeditions from 1950 to 2019, as shown in fig6 for the total market sizing. We used an auto-ets model and a simple exponential smoothing model for the forecast. The graph suggests that the market size will remain at the current level.

**Insurance Companies**

From the insurance companies’ perspective, we built two prediction models to predict the outcome and highest point of one expedition, respectively. For the outcome prediction model, we adopted three models, which are logistic regression, gradient boost machine, and neural network. We believe the neural network model as the optimal model, as this model has the highest accuracy number.

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy Measure | Logistic Regression | Gradient Boost Machine | Neural Networks |
| Accuracy | 0.609 | 0.969 | 0.984 |
| Specificity | 0.997 | 0.991 | 0.984 |
| Sensitivity | 0.229 | 0.948 | 0.985 |

For the highest height prediction model, we adopted linear regression, gradient boost machine, and neural network. We consider the neural network model as the optimal model, as it has the lowest RMSE.

|  |  |
| --- | --- |
| Model | RMSE |
| Linear Regression | 1221 |
| Gradient Boost Machine | 898 |
| Neural Networks | 892 |

**Conclusion**

# In recent years, with the development of technology and the economy, more and more middle-aged and elderly want to participate and have the experience of climbing Himalayan. The customers portraits for this group of people are financial independence, relatively preserved, and value health and safety over adventure. Tour companies can focus on this group of people, by conducting more specific research and offering luxury climbing packages for them with feature of advanced equipment and experienced tour guide at a cost of $40,000 to $60,000.

# At the same time, insurance, which is one of the necessities of climbers' journeys, has also spawned a larger market. We have successfully built two models predicting the outcome and the highest point of expedition using a neural network. The insurance company can utilize our findings to achieve higher efficiency, by adding services such as helicopter evacuation, oxygen cylinder delivery and trip interruption coverage accordingly.

**Recommendation for further research**

For the trend analysis part, we only focus on the citizens and peaks of highest occurrence, which involve limited parameters. For further analysis, we can expand the scope of our research by adding more variables and digging deeper. By doing that, we can present a more comprehensive model that includes more variables.

For the prediction model, we didn’t include the weather variables while they seem like obvious factor. In order to provide greater effectiveness of the model, we can add this factor in the future by cooperating weather through third-party API using date and location variables in current datasets. We could also incorporate the geography effects on our model to enhance our prediction accuracy in the future.

# 

# **Appendix：**

Data Description Table

|  |  |
| --- | --- |
| **Variable Name** | **Variable Description** |
| expid | Expedition ID |
| year | Year |
| season | Season |
| route1 | The climbing route summary |
| nation | The principal nationality of the team |
| leaders | Leader’s name for each name |
| sponsor | The sponsor party of the expedition operation |
| camps | Number of high camps above the base camp |
| rope | Amount of fixed rope (meters) |
| totmembers | Number of members in the team |
| smtmembers | Number of members on summit |
| tothired | Number of hired personnel above base camp |
| smthired | Number of hired personnel on summit |
| o2used | Whether oxygen was used during the whole operation |
| o2climb | Whether oxygen was used during climbing |
| o2sleep | Whether oxygen was used during sleeping |
| o2medical | Whether oxygen was used for medical purposes |
| comrte | Whether the route taken was the commercial route |
| stdrte | Whether the route taken was the standard route |
| sex\_ratio | The proportion of males on the whole team |
| average\_age | The average age of the team |
| median\_age | The median age of the team |
| leader\_ratio | The proportion of leaders on the team |
| deputy\_ratio | The proportion of deputies on the team |
| staff\_ratio | The proportion of staff on the team |
| sherpa\_ratio | The proportion of sherpa on the team |
| tibetan\_ratio | The proportion of tibetan on the team |
| speed | The proportion of members who speed climbed on the team |
| o2used\_ratio | The proportion of members who used oxygen during the whole operation |
| o2climb\_ratio | The proportion of members who used oxygen during climbing |
| o2decent\_ratio | The proportion of members who used oxygen during descending |
| o2sleep\_ratio | The proportion of members who used oxygen during sleeping |
| o2medical\_ratio | The proportion of members who used oxygen for medical purposes |
| disputed | Whether the claim of success is either unverified, has been disputed by another party, or is of a controversial style. |
| claimed | Whether the claim of success has been disproved or is not generally recognized by the mountaineering community |
| termreason | Reason of termination |
| success1 | Whether the operation succeeded |
| smtdays | Number of days it took to the summit or reach the highest point |
| highpoint | The highest point team reached |
| mdeaths | Number of member deaths |
| hdeaths | Number of hired personnel deaths |
| success\_ratio | The proportion of members who were defined success on the team |
| death\_ratio | The proportion of members who died in the operation |

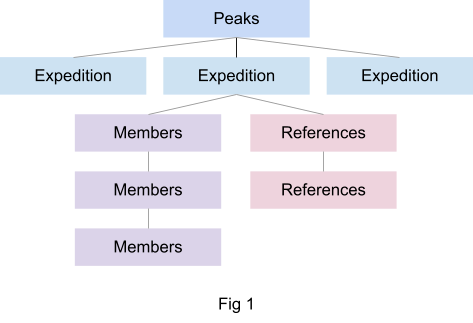


Fig 1

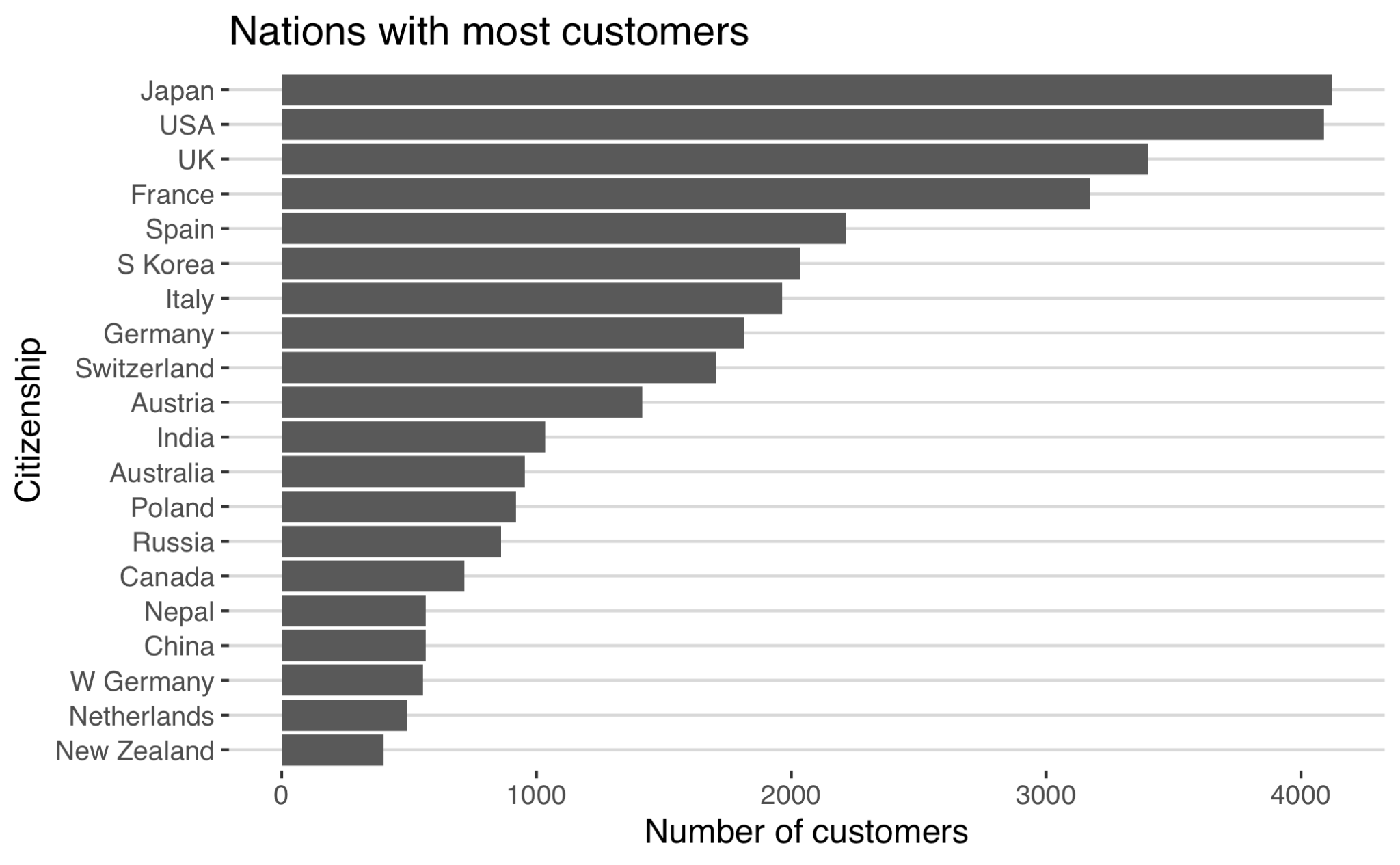


Fig 2

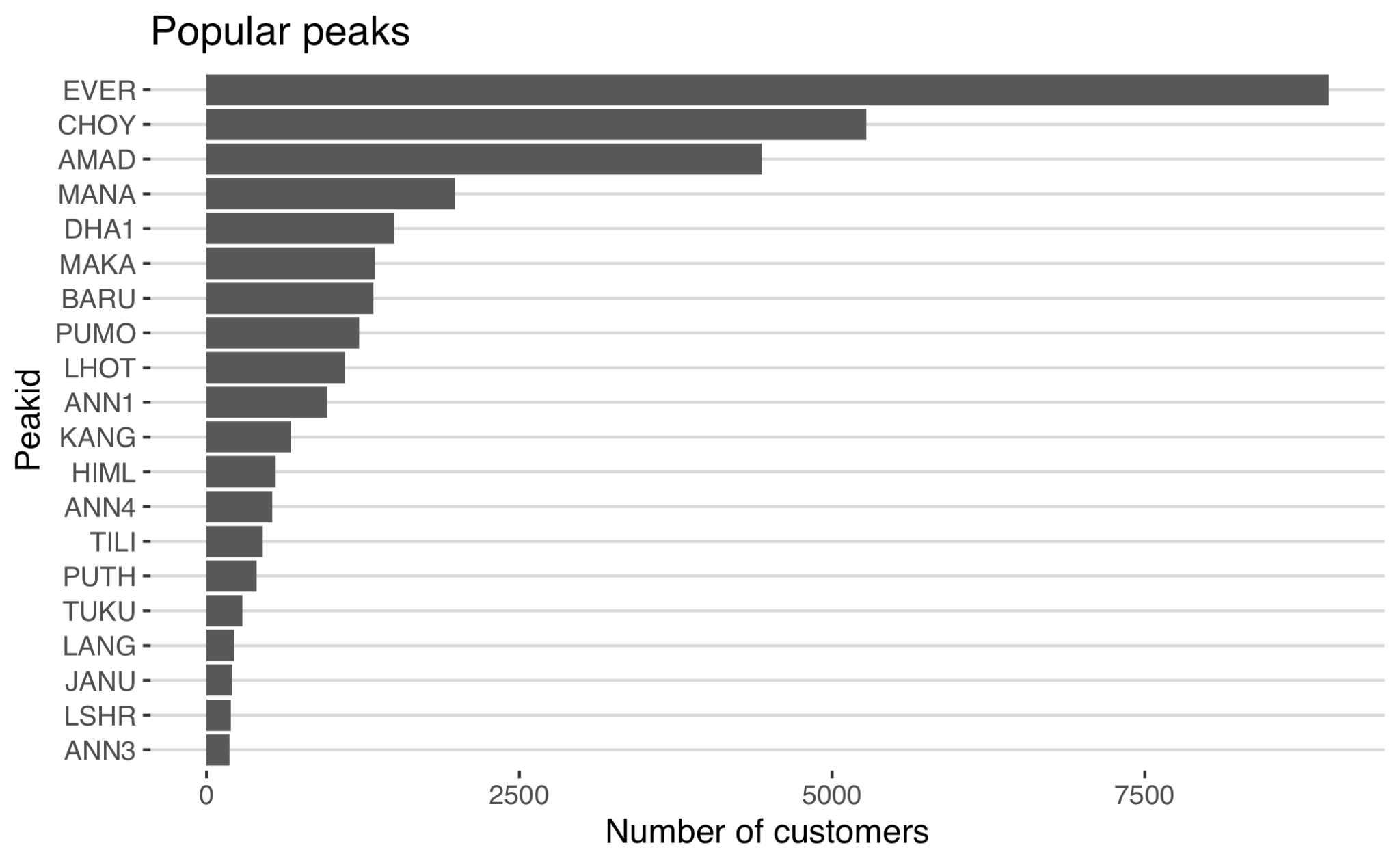


Fig 3

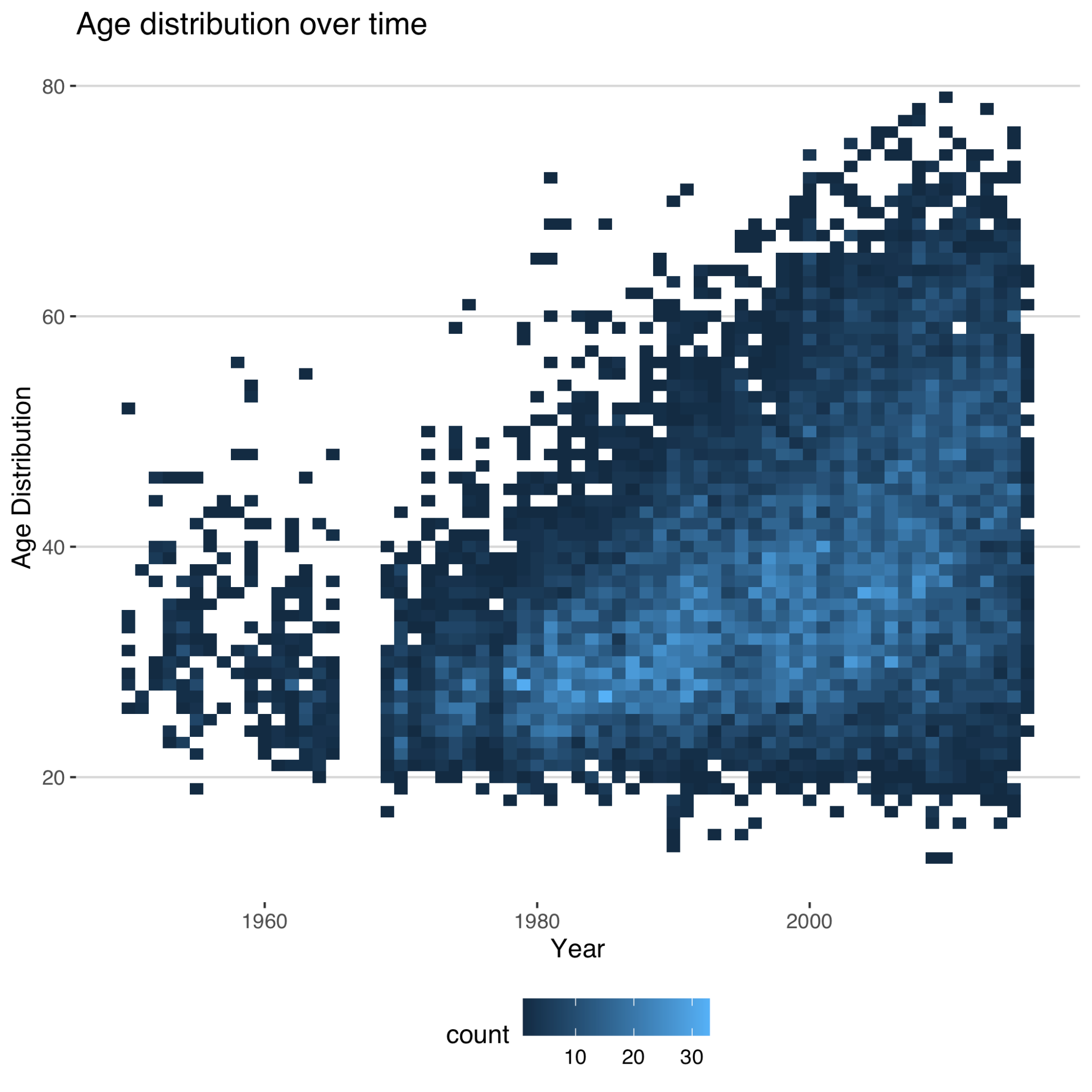


Fig 4

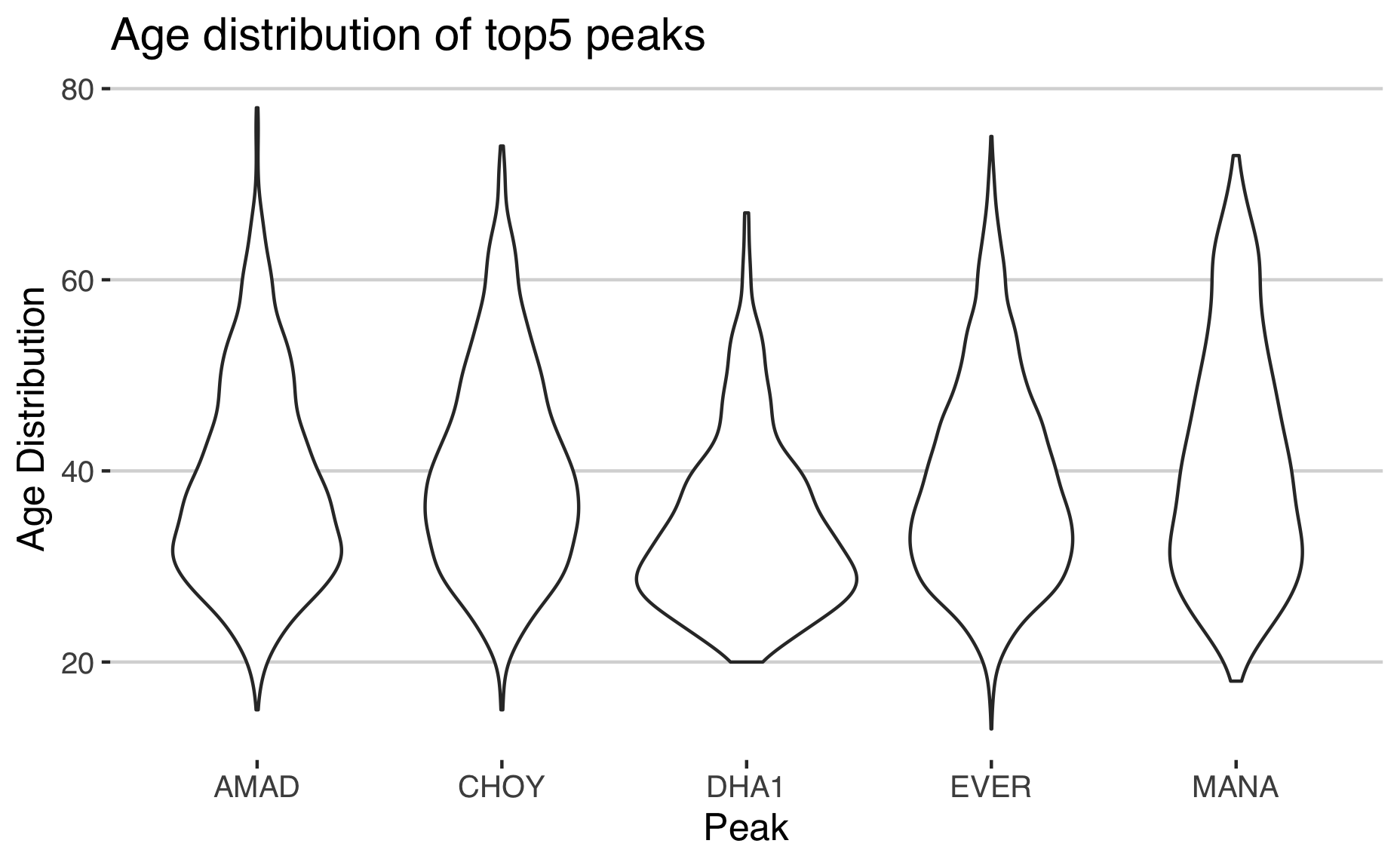


Fig 5

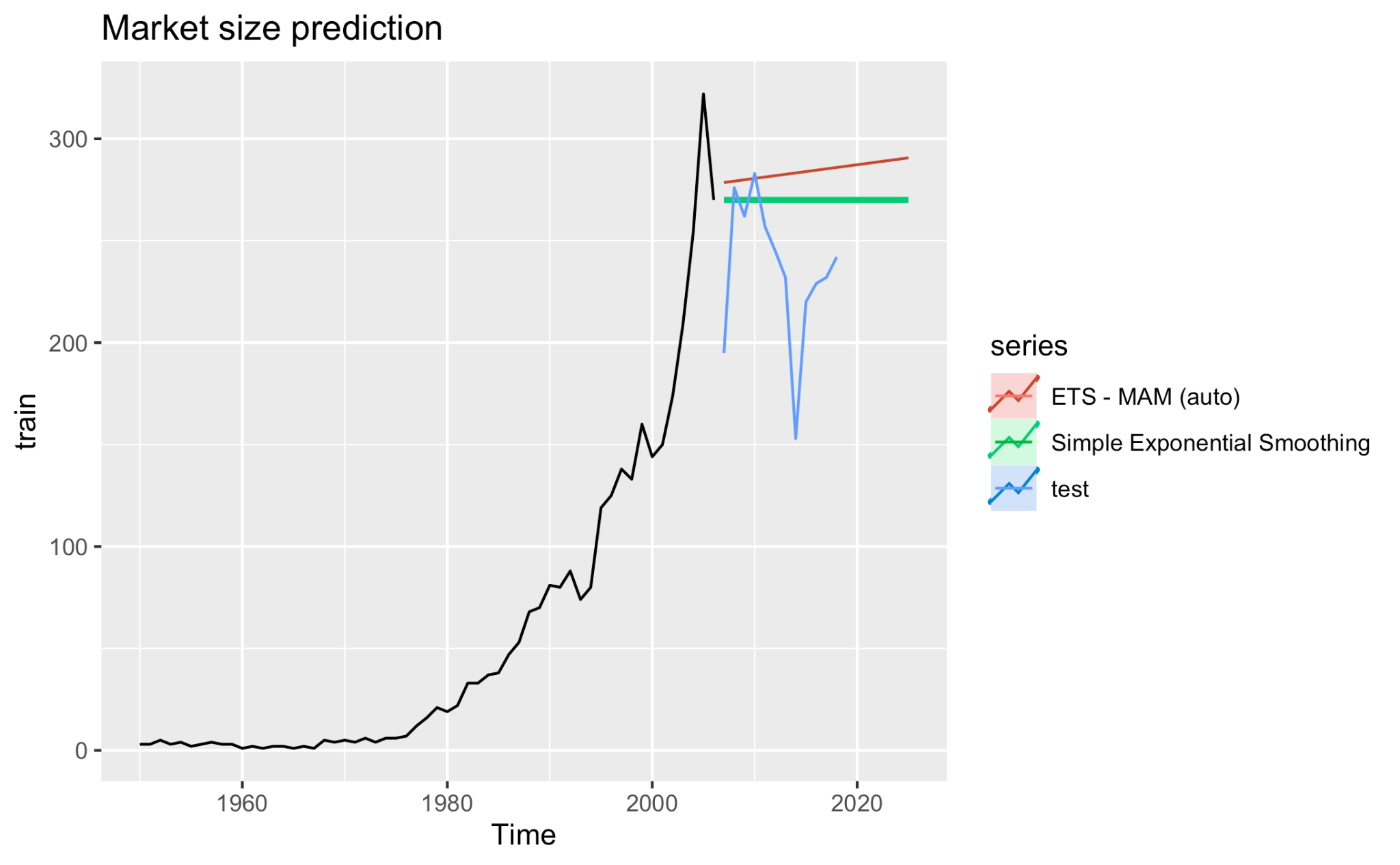


Fig 6

**Reference**

Salisbury, R., Hawley, E., & Bierling, B. (2011). *The Himalaya by the numbers: A statistical analysis of mountaineering in the Nepal himalaya*. Vajra Publications.

Huey, R. B., Carroll, C., Salisbury, R., & Wang, J.-L. (2020). Mountaineers on Mount Everest: Effects of age, sex, experience, and crowding on rates of success and death. *PLOS ONE*, *15*(8). https://doi.org/10.1371/journal.pone.0236919

Li, G. (n.d.). *Everest climbing economic account: starting at 460,000, every meter is "money"*. 21jingji. Retrieved March 3, 2022, from https://m.21jingji.com/article/20190531/69d2f3b09356210080c54ba955051eb4.html