SYSTEM DOCUMENTATION

XYZ Corporation

Team : Pathfinders

Members:

- Amandeep Singh Arora
- Shawn Casaus
- Robert Pitchford
- Ravi Verukonda

Table of Contents

Introduction

This document provides a detailed overview of XYZ Corporation's project to assist UVW College in its marketing efforts to bolster enrollment and determine other important information developing marketing profiles on people based on the data supplied by the United States Census Bureau. As part of this project, visualizations are developed which help in identifying the key factors to support the development of the model/application

Business Objective

The business objective is to predict the salary range that helps UVW College to bolster enrollment and achieve its target.

Roles and responsibilities

The primary roles in this project are product owner, team members and stakeholders UVW College and XYZ Corporation.

The responsibilities per role are as below

Product Owner:

- Work with UVW College and gather the requirements
- Finalize the features that need to be developed and delivered
- Defining the acceptance criteria of the user stories
- Prioritizing the stories or features
- Managing the project expectations, timelines and dependencies
- Evaluating the project progress at each iteration by participating in user stories' demonstrations

Team Members

- Understanding the requirements and feasibility of the implementation
- Analyze the data supplied by the United States Census Bureau
- Identify the tools and technologies that can be used to develop
- Develop visualizations that can be used for building profiles
- Deliver the expected artifacts like System Report Documentation, Visualizations and Executive Report
- Work in an Agile model and deliver user stories as prioritized by Product Owner
- Validate the developed visualizations

Stakeholders

- UVW College :
  - Responsible for participating in requirements gathering with XYZ Corporation
  - Responsible for providing clear requirements for developing the model
  - Responsible for providing the acceptance criteria
- XYZ Corporation
  - Responsible for creating the key factors to develop marketing profiles
  - Responsible for development of the application to predict the income of an individual based on different values of the input parameters so that UVW College can tailor their marketing efforts when reaching out to the individuals

Team goals

The important goals of the team are:

1. Well-groomed user stories with acceptance criteria and effort estimation
2. As part of the exploratory analysis, identify common patterns and outliers
3. Compare the data for similarities and differences for filtering
4. Mapping the data as required
5. Rendering the data using different visualizations
6. Finalize the visualizations which help in identifying the key factors to develop marketing profiles
7. Develop the Executive report with the selected visualizations

Assumptions

The assumptions made for this project are

1. Salary prediction is good enough for UVW College to achieve its enrollment target
2. UVW College will provide the expected accuracy of the model to be developed
3. The model developed based on the data provided will reach UVW College's expected accuracy
4. Based on the feedback received from UVW College after a trial run of the model, updates are made in the model as required to improve the accuracy
5. The provided data set is a good fit for developing the model/application
6. The application team shall be able to understand the visualizations developed and come up with model/application to predict the salary range

User Stories

The user stories identified as part of the project are as below

Use Case Name	Predict Salary (Feature)
Description	Predict the salary range of a given individual
Primary Actor	UVW College
Basic Flow	To bolster recruitment, UVW College would like to be able to predict the salary range for any given individual given a key set of variables describing that individual. The salary should be classified as either <=$50k or >$50k.

Use Case Name	Determine Factors
Description	Determine key variables for input to the model
Primary Actor	XYZ Corporation
Basic Flow	Develop marketing profiles using data supplied by the United States Census Bureau, and you will be focusing on $50,000 as a key number for salary. Via data visualization, determine the most important predictive parameters to be included in the model.

Use Case Name	Develop Model
Description	Develop a model capable of predicting salary
Primary Actor	XYZ Corporation
Basic Flow	Develop a model that can predict the salary range [<=$50k, >$50k] of an individual from a set of key identifying parameters. Accuracy is determined by the F1-score of the model which must be >0.95. The set of parameters is provided by the Determine Factors Use Case.

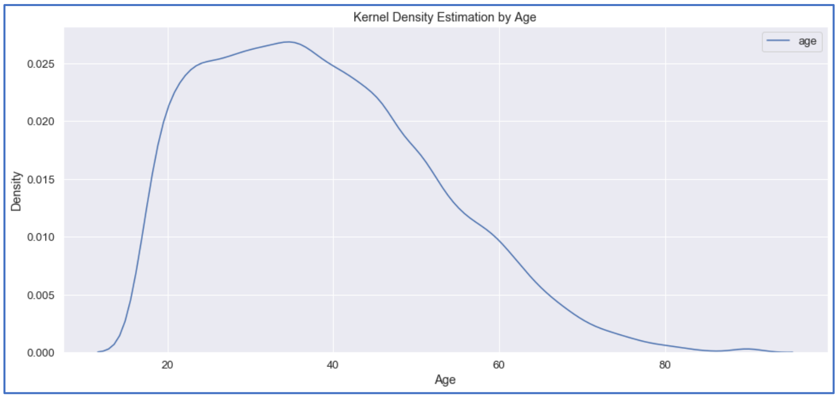
Visualizations

To identify the key factors to predict the salary range, firstly understand the data by different visualizations, evaluate which factors are dominant over other factors and come to a conclusion.

Understand Data & Evaluate

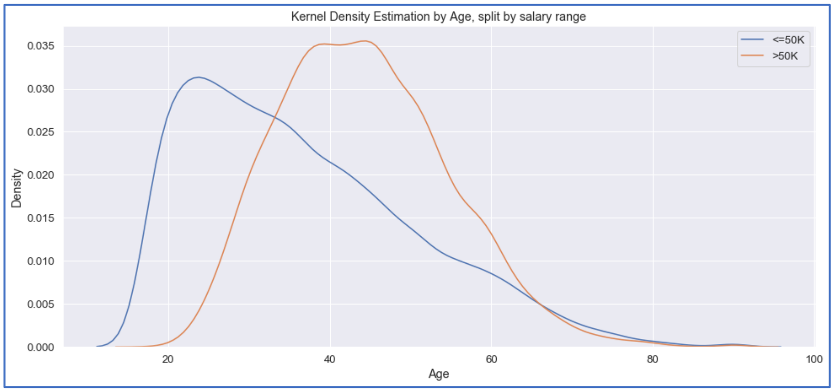
As part of the understanding data and evaluation, go over the dimensions (columns) part of the data set and develop visualizations for every dimension and identify the ones which impact the salary range mostly.

The line graph in **Fig(1)** depicts the distribution of the age groups with mean 38.6, median 37, lower quartile 28, upper quartile 48, min 17 and max 90.



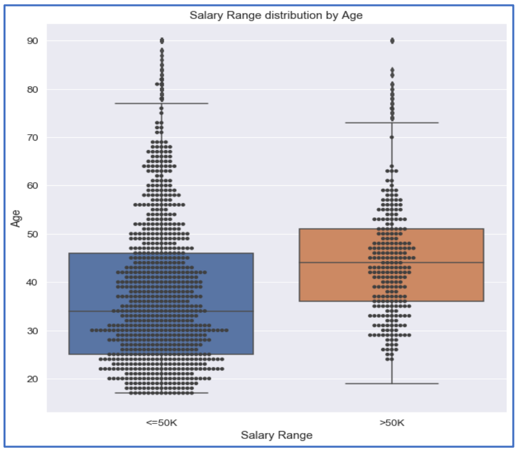
Fig(1)

The line graph in **Fig(2)** depicts the distribution of the age groups in the data set split by salary. We can see that younger people have a lower salary range.



Fig(2)

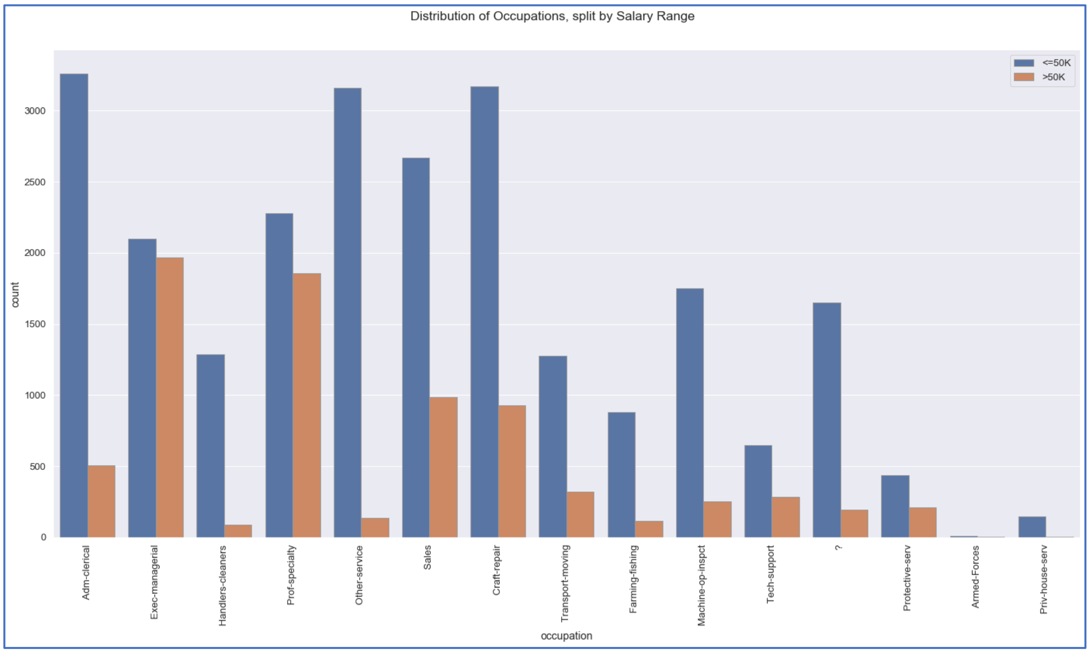
The Box and Whisker Plot as shown in Fig(3) , depicts that the age group 36 to 51 has a salary range mostly >50K with a median 44 and the age group 25 to 46 has a salary range mostly <=50K with median 34.



Fig(3)

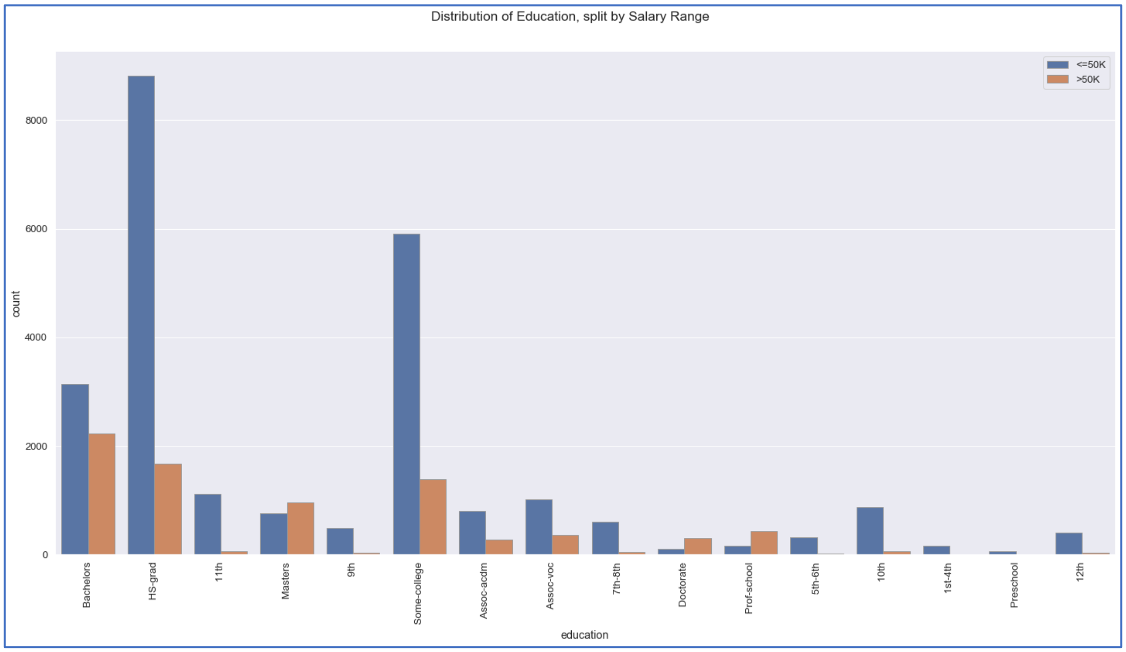
Fig(4) is a bar chart for Occupation against salary and split by salary ranges.

The salary range <=50K is dominant for most of the occupations except Exec-managerial and Prof-specialty



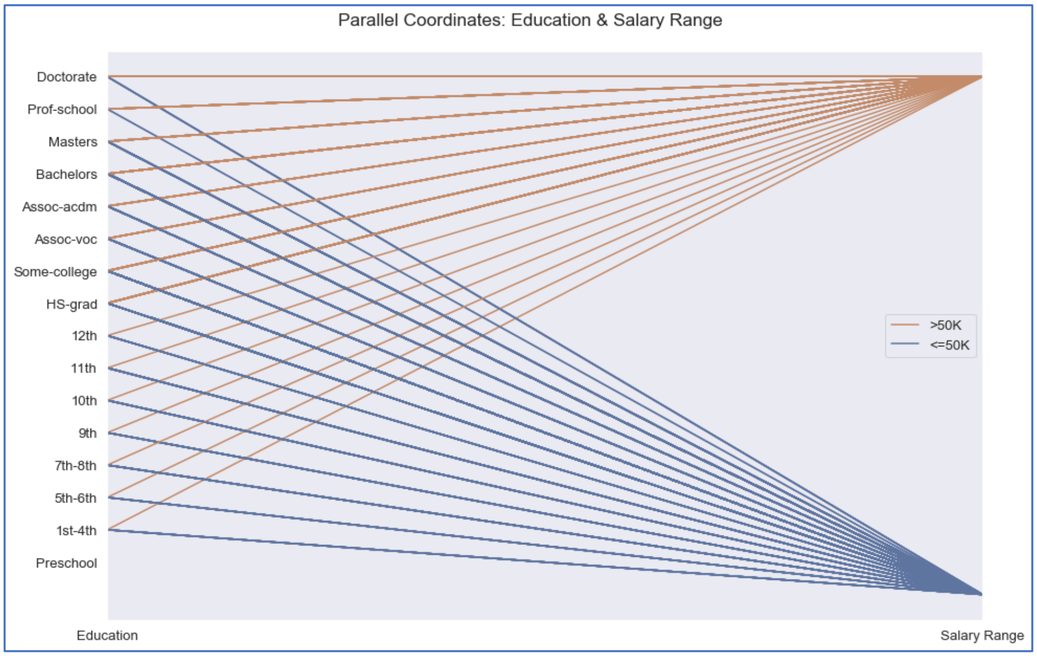
Fig(4)

Fig(5) is a bar chart for Education against salary split by salary ranges.



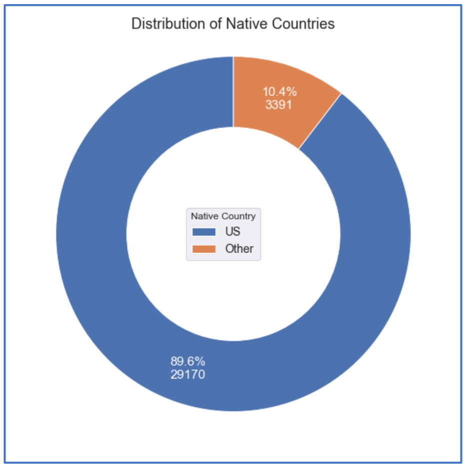
Fig(5)

Fig(6) is a parallel coordinates plot using education num and salary ranges. As it is evenly distributed and most of the education numbers are mapped to both salary ranges, we cannot consider this as a good candidate.

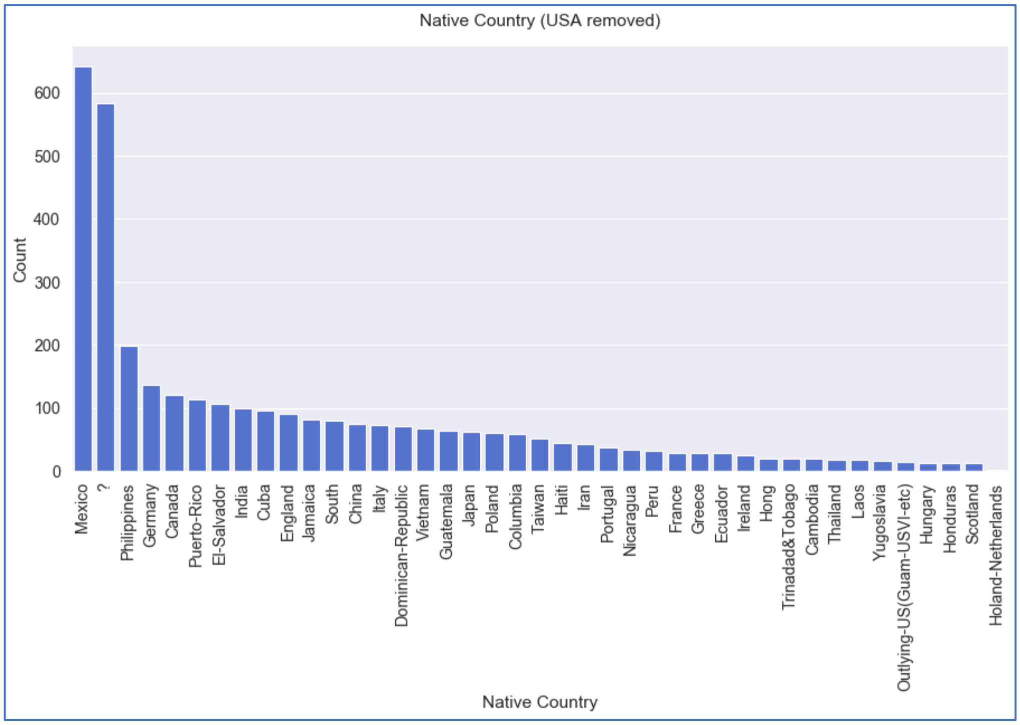


Fig(6)

Fig(7) shows that the most of the responders are from the United States and Fig(8) is a bar chart showing the number of respondents per native-country except the United States and we see that unknown '?' is the next dominant after Mexico making the native-country not a good candidate to consider for the business objective.

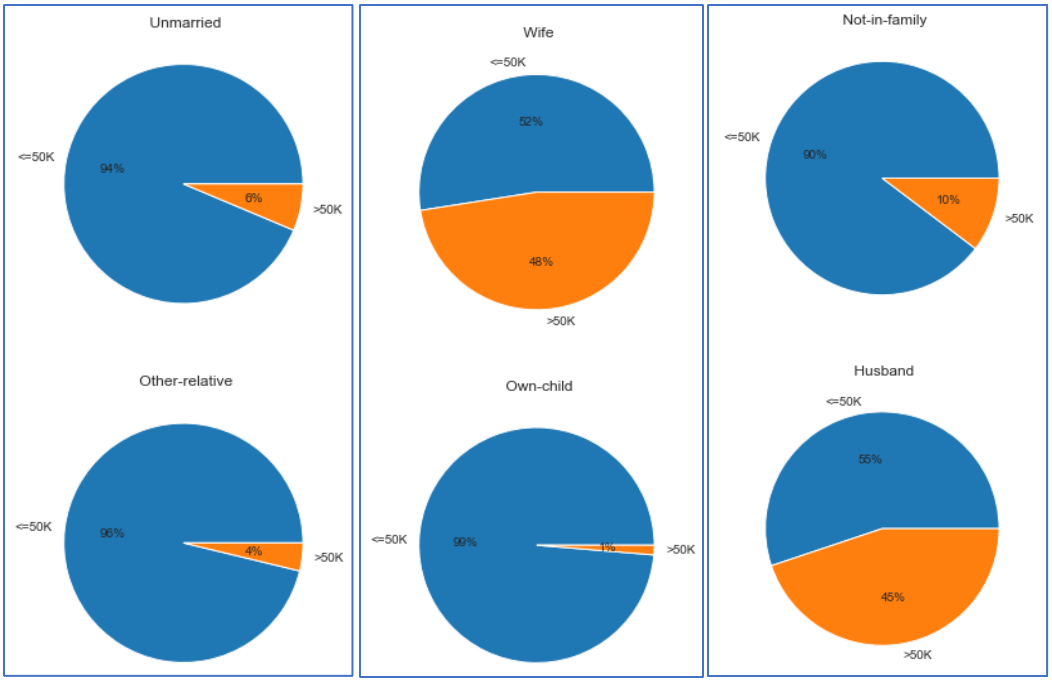


Fig(7)



Fig(8)

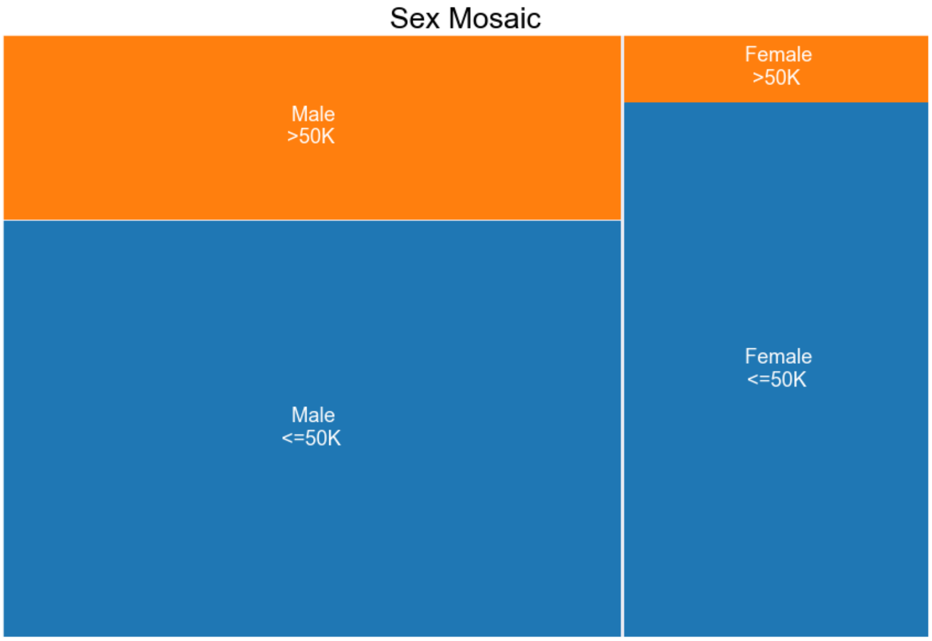
Fig(9) is a set of pie charts depicting salary ranges across relationships and we see that the Wife and Husband are with most responders with salary range >50K and others are mostly <=50K



Set of Pie Charts – Salary Range Distribution per relationship

Fig(9)

This Mosaic plot from Fig (10) depicts the Sex distribution and we see that most of the salary ranges are <=50K and Females are more likely to earn <=50K



Fig(10)

Conclusion

Based on the visualizations, we see that the occupation, relationship and sex are the key dimensions that clearly distinguish the salary ranges of the respondents. The other dimensions like native-country are not contributing to distinguish the salary range as there is missing information and a major portion of the data has native-country has the US with mixed salary ranges.

The Age dimension can be used as the first rule like any person with age less than 20 years and greater than 75 years will have a salary range <=50K. The respondents with relationships unmarried, own a child, not in family and others make <=50K.

The respondents with occupations except Exec-managerial and Prof-specialty make <=50K.

Questions

Below are the problems during the project progression and described solutions that are followed while developing the visualizations.

1. There are column values with '?', do we need to exclude the rows with such column values

   Its good to consider '?' column values as there is a possibility that the United States Census Bureau may not have all the fields on a profile.

2. What all visualizations need to be developed?

   Bar Chart to understand the distribution of the data

   Box and Whisker plot to understand the age distribution with outliers

   Pie charts & Mosaic plot to identify which variant of dimension has distinguishable salary ranges

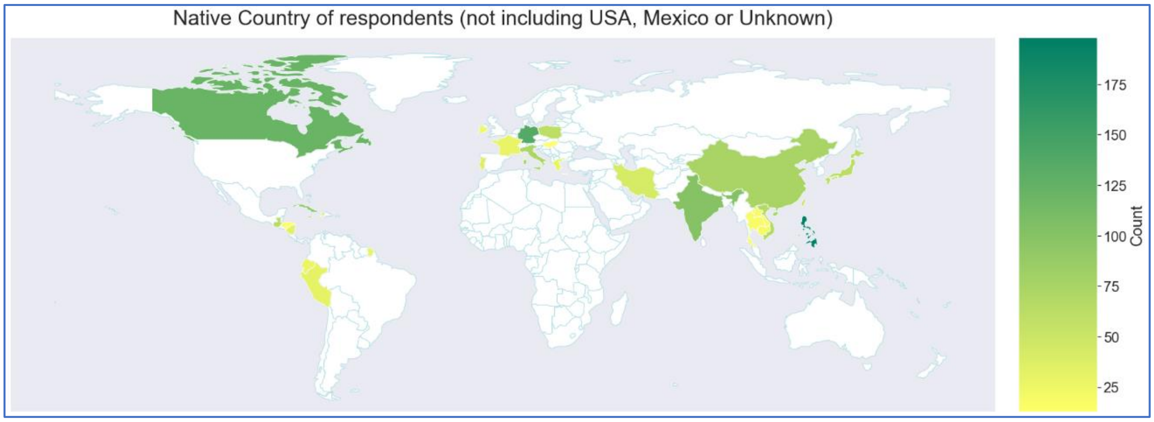
   Parallel co-ordinate is useful to identify the relation between two dimensions

3. How to choose a visualization as a prime candidate for the application/model?

   Visualizations with distinguishable salary ranges need to be selected

Future Plan

Based on the data set, we observed from Fig (11) that there is a potential to identify the foreign profiles for marketing to bolster foreign students' enrollment


Native Country of respondents (not including USA, Mexico or Unknown)
175
150
125
100
75
50
25
Count

Fig(11)