System Documentation Report

Team 35

Updated: 06/28/2020

Team Members

- Amit Sharma
- Eric Peters
- Oana Almasan
- Claudio Rodriguez Rodriguez

Table of Contents

Course Project Introduction

We are XYZ Corporation, and we use data to develop marketing profiles on people. We provide these profiles to numerous companies for marketing purposes. The following document is a joint effort with UVW College, a local college looking to bolster enrollment. UVW has chosen a salary as a key demographic to determine the criteria for marketing its degree programs.
In the document, we will observe different marketing profiles using data supplied by the United States Census Bureau, and we will be focusing on $50,000 as a critical number for salary.
For example, if the data show that the majority of individuals making less than $50,000 is under 34 years old, male, single, and has a high school diploma, the college can market with tuition amounts, program concentrations, and even ground or online programs appropriate to this demographic.
To achieve its enrollment target, the marketing team at UVW would like to develop an application to find the factors that determine the individual's income. We create Marketing Profiles in the form of User Personas for UVW, so they use them to develop their model. These profiles will empower UVW to predict the income of an application based on different input parameters to tailor their marketing efforts when reaching individuals.

Roles and Responsibilities

Product Owner

UVW College

Team Members

- Amit Sharma
- Eric Peters
- Oana Almasan
- Claudio Rodriguez Rodriguez

Stakeholders

- UVW College
- XYZ Corporation

Team Goals and Business Objectives

1. Develop Marketing Profiles for UVW College so that they can develop marketing materials that focus on a specific user persona.

Assumptions

Provide a brief explanation of the strategic aim of your actions. Why are you building the product? How do your actions affect product development and align with the company's goals?

- Planned offerings include Degree programs, speciality vocational courses, financial aids and grants to minorities.
- The College board is interested in building profiles based on various parameters like gender, ethnicity, race, occupation, age, salary etc.
- Assuming Hong stands for Hong Kong because the dataset is from 1994 when Hong Kong was a UK Colony.
- Assuming South is South Korea instead of South Africa.
- Assuming Columbia is Colombia.
- Some-College correlated

User Stories

List or link user stories that are required for the project. A user story is a document written from the point of view of a person using your software product. The user story is a short description of customer actions and results they want to achieve.

1. As a College I want to see the distribution of Education per Country.
2. As Management of UVW College, we want to know the influential factors to consider that affect a person's Salary.
3. As Management of UVW College, we want to see the years of education by Salary and Ethnicity to figure out to whom we should market our multiple Educational Programs.
4. As Management we want to find out Education Opportunities depending on the Job performed.
5. As Management we want to find out the critical age to target our Programs.
6. As Management we want to find out how individuals are distributed based on their family relationship.
7. As Management of UVW College, we want to find out the target audience to offer the education programs based on hours-per-week and salary.
8. Management is considering a program to people in different marital status and wants to know if it is feasible.
9. As Management of UVW College I want to observe the distribution of salary per country to propose help to minority groups.
10. As management of UVW College I need to know the most important traits of prospective graduate and, respectively undergraduate students

Visualizations

Link the visualizations created to the business objective. Explain what each visualization tells you and why you choose to create the visualization.

As a College I want to see the distribution of Education per Country.

- **User Persona**: Multiple User Personas shown
- **Requirement**: As a College, the management wants to see the general distribution of Education per country to understand the target base. This will help the organization to understand the requirement and reach on the broader level.
- **Answer:** We could notice that people from countries like Trinadad&Tobago, Guatemala, Philippines and Puerto-Rico has lower education levels than other countries. Also, we could see that countries like India do have the highest level of education recorded by the census.
- **User Scenario**:
  - The graph below shows the most common attained Education Level per Country.
  - This will help the college management to target their materials and programs to the specific demographic.
- **Visualizations**:
  - Refer Appendix(I).

As Management of UVW College, we want to know the influential factors to consider that affect a person's Salary.

- **User Persona**: Multiple User Personas shown
- **Requirement**: What are the most influential factors to consider that affect a person's salary?
- **Answer:** We could notice that age, relationship, education-num, capital gain, marital-status, occupation and hours-per-week are some important features that impact or influence the outcome of the model which is salary.
- **User Scenario**:
  - We have implemented CatBoostClassifier, to predict and classify the outcome of the model designed with the help of given data.
  - We want to understand the most important factors that influence the outcome, which is a person's salary. The outcome could be >50K(labeled as 1) or <=50K(labeled as 0).
  - We have used a summary plot to aggregate the SHAP values for all the features and all samples in the selected set.
  - The values are then sorted and shown in the order of decreasing importance.
  - If we deep dive and try to look at the other summary plot, we can further see the positive and negative relationship between the feature predicate and the target salary.
  - Let's consider age; it could be seen that age plays a very important role and lower the

value of age; the model output was 0(negative SHAP Value; salary <=50K). It totally makes sense: if the age is low, the salary would also be low as the education level and experience is low.

- With increasing age, you get the opportunity to increase your education level and experience, thus increasing the opportunity to earn salary >50K.

- **Visualizations**:
  - Refer Appendix(II).

As Management of UVW College, we want to see the years of education by Salary and Ethnicity to figure out to whom we should market our multiple Educational Programs.

- **User Persona**: Multiple User Personas shown
- **Requirement**: How to market multiple programs based on ethnicity, grouped by salary and education.
- **Answer:** We could notice that Amer-Indian-Eskimo Males and Females have a high density of finished High School but not a lot of them finish Bachelors Programs. Similarly, Asian-Pac-Islander race seems more interested in Master's and Doctorate Programs, and Women of most races except Amer-Indian-Eskimo show success in Doctorate programs.
- **User Scenario**:
  - UVW College marketing department wants to correlate years of education with Salary
  - Helps to know to which Nationality they should market their Bachelors Programs.
  - Helps discover possible Government programs for minority groups
  - Allows for more specific targeting of gender groups.
- **Visualizations**:
  - Refer Appendix(III).

As Management we want to find out Education Opportunities depending on the Job performed.

- **User Persona**: Multiple User Personas shown
- **Requirement**: What are different education opportunities that we can provide based on jobs performed?
- **Answer:** We could notice that people with jobs like Craft-Repair and tech Support with lower years of education can be offered with a profession-based specialty course to improve their skills and earning.
- **User Scenario**:
  - We observe that people in the Machine-op-inspct, Craft-Repair, Tech-Support, Adm-clerical Job Industry are excellent candidates for offering a bachelor's Program.
  - We can also observe that people in Exec-Managerial, Prof-specialty are excellent candidates for master's and PhD Programs.
- **Visualizations**:

- Refer Appendix(IV).

As Management we want to find out the critical age to target our Programs.

- **User Persona**: Multiple User Personas shown
- **Requirement**: What's the critical age for the college offerings?
- **Answer:** From the plot, we could see various age groups for different education degree programs. The user scenario mentions few of those.
- **User Scenario**:
  - We observe the biggest increase in Bachelor programs starting around age 17 to 18, which is consistent with the common age High School ends, but we see High School maintaining popularity over the years. Finally, we observe an increase in seriousness of bachelor's Programs for Women in their early twenties.
  - The highest earning education levels see a higher peak in education years at later stages, and it is more common to find an Exec-managerial role with a finished Bachelors.
  - Masters Programs interests remain low statistically but consistent across young adults.
- **Visualizations**:
  - Refer Appendix(V).

As Management we want to find out how individuals are distributed based on their family relationship.

- **User Persona**: Multiple User Personas shown
- **Requirement:** What's the distribution of individuals based on relationships?
- **Answer:** From the plots, it is evident that Relationship is one of the major predictors for determining the income.
- **User Scenario**:
  - We can see that the data may indicate that some of the reported salaries were "household" salaries, rather than individual salaries
  - Married individuals (husbands and wives) appear to have a much higher proportion of >50K earners
  - Individuals not in a family appear to be more likely to have a higher number of years in education
  - Based on the SHAP analysis, family role is the strongest predictor of salary, likely due to the above logic of 'household income' vs 'individual income'
- **Visualizations**:
  - Refer Appendix(VI).

As Management of UVW College, we want to find out the target audience to offer the education programs based on hours-per-week and salary.

- **User Persona**: Multiple User Personas shown
- **Requirement:** What is my target audience based on the hours-per-week and salary?
- **Answer:** We can target people with lower education level working greater hours per week but are earning income less than 50K.
- **User Scenario**:
    - We could observe that the majority of people with lower education level, have to put in more hours per week in order to be able to earn more than 50K.
    - At the same time, we could also observe that people with a number of education years 14,15,16(Masters,Doctorate,Prof-school), even if working for 20-40 hours a week could end up earning more than 50K.
    - We can also observe that there is no correlation between working more hours and earning more salary. While it would be safe to assume that "less time spent on education" might be correlated, there is no data to back this statement.
- **Visualizations**:
    - Refer Appendix(VII).

Management is considering a program to people in different marital status and wants to know if it is feasible.

- **User Persona**: Differentiated by Marital Status with a focus on Married and Separated Users.
- **Requirement:** Can we sell profiles based on Marital status.
- **Answer:** Yes, the observation indicates some correlation. Refer to the user scenarios.
- **User Scenario**:
    - Our analysis indicates that Married with a Civilian Spouse has the highest salary distribution.
    - Single and Separated groups are in the lower salary spectrum. Still, it's complex to drive conclusions from this statement because the lower salary bracket might be the reason for them being in this group. Without more data to support it no conclusions can be derived.
    - There are multiple Married Groups, and they all seem to earn more money. Still, it is not enough of a relationship to conclude that Married leads to more salary due to the substantial number of outliers in every group.
- **Visualizations**:
    - Refer Appendix(VIII).

As Management of UVW College I want to observe the distribution of salary per country to propose help to minority groups.

- **User Persona**: A minority group would be a group of people that has a disadvantage compared to other groups. They would be in the same age groups as other groups looking for College Education (18-30 years old).
- **Requirement:** UVW College is providing a program partnered with Government Grants and Local Business Scholarships to assist minority groups.
- **Answer:** There are many minority groups that would benefit from a Government supported program, and paired with the relationship that higher education has with higher salary then helping them achieve a Bachelors would prove as helpful to them.
- **Visualizations**:
  - Refer Appendix(IX).

As management of UVW College I need to know the most important traits of prospective graduate and, respectively undergraduate students

- **User Persona:** multiple user personas
- **Requirement:** the client wants to develop new programs (both degree and non-degree) targeting undergraduates and graduates including adults in continuing education and adults returning to organized education.
- **Answer:** prospective undergraduate - not married, < 30 work-hours/week, no investments, age under 25, non-managerial/low skill job, most probably woman, income under 50k; prospective grad: married, works 40+ hrs/week, notable investment activity, age >25, holds a BA, has a management/high skill job, most probably man, income >50k/year.
- **User Scenario:** based on a binary classification prediction model (XGBoost - test accuracy 86%), SHAP - SHapley Additive exPlanations was used to explain the results:
  - Positive and negative factors affecting individual income **-** a significantly higher weight of the first seven features:

    1. **marital status** --> married individuals are more likely to have an income over 50k
    2. **work-hours/week** --> the higher the better (optimal between 40 and 60)
    3. **capital gain (or loss in other cases)** --> the higher the better (any activity in this area is an indicator of better financial status)
    4. **age** --> the higher the better
    5. **education** --> lack of BA degree correlates with an income lower than 50k.
    6. **occupation** --> having a management position or a professionally specialized job is an indicator of higher income.
    7. **gender** --> the income gender bias is confirmed, men earning significantly higher than women

- **Visualizations**:
    - refer to Appendix (X)

Questions

As teams solve problems during the project progression, many questions inevitably arise. Report the questions your team had and the solutions you implemented.

- Country labeled '?' had a non-trivial amount of entries, was it null?
- What is South and Columbia in the Native Countries?
- Are the Army numbers due to lack of information? Is Army education not counted in the census?
- There's ~22k Male samples versus ~11 Female Samples yet the distribution of Sex across that country tends to be close to 50% each, is the data reliable?
- There is no data for Africa.
- Should we drop the fnlwgt variable? (as '[...]the CPS sample is actually a collection of 51 state samples, each with its own probability of selection, the statement only applies within state');
- Was the annual income recorded as individual or as household income?

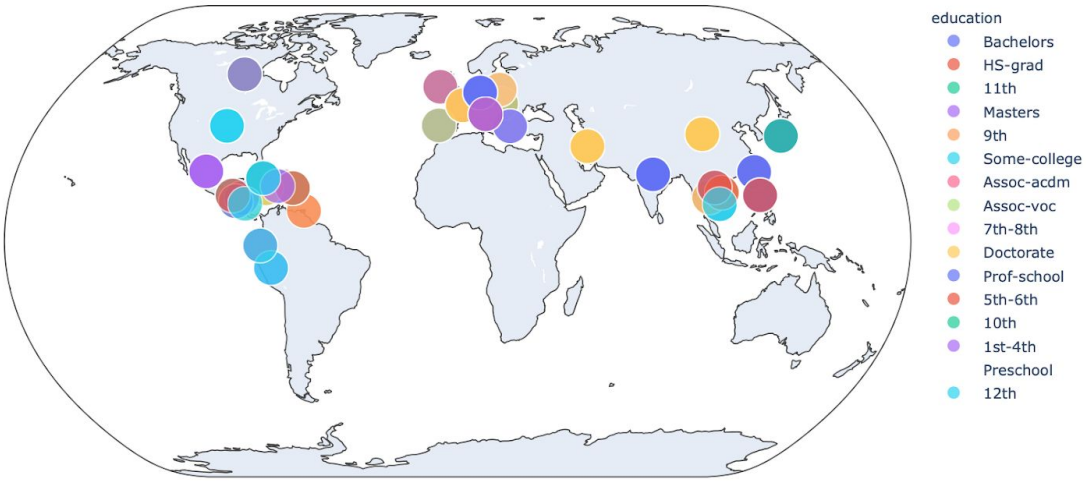
Not doing

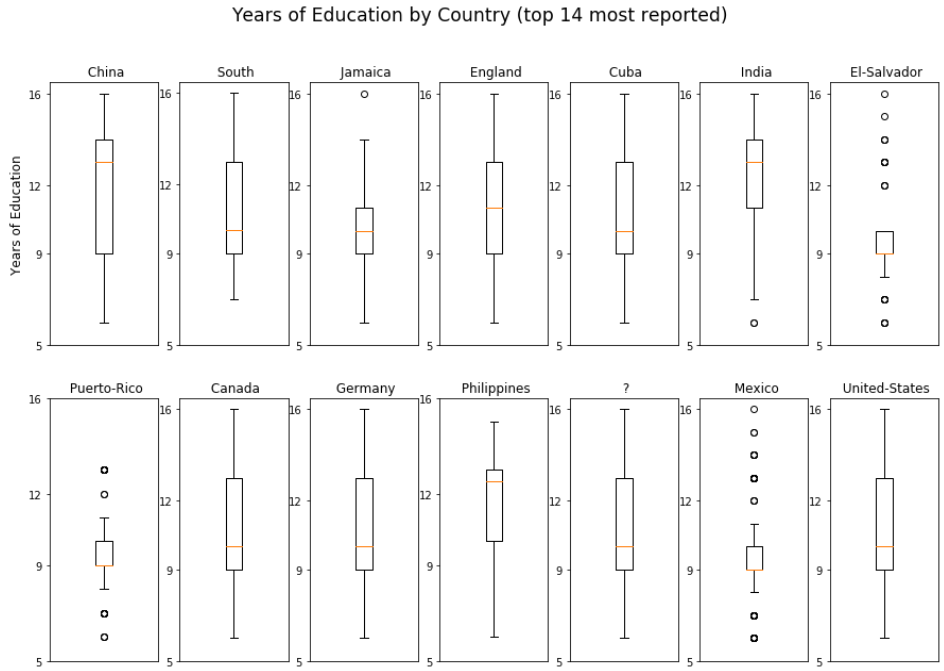
List the things that you are not doing now but plan on doing in the future. Such a list will help the developers prioritize features.

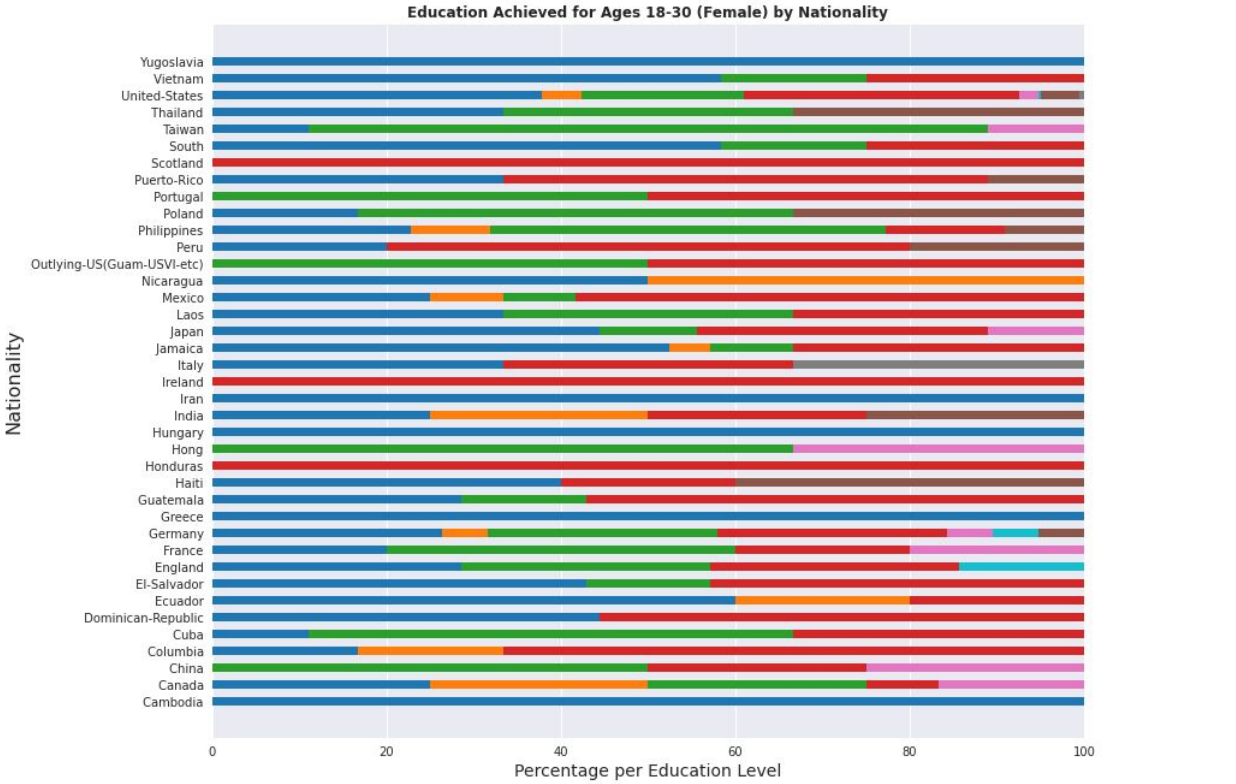
- Models used to segregate variables are not deliverables for UVW College.
- Models were not used for prediction but are used to discover the importance of each attribute in the dataset.
- Conclusions from Models used are not to be used as conclusions of Customers' asks.

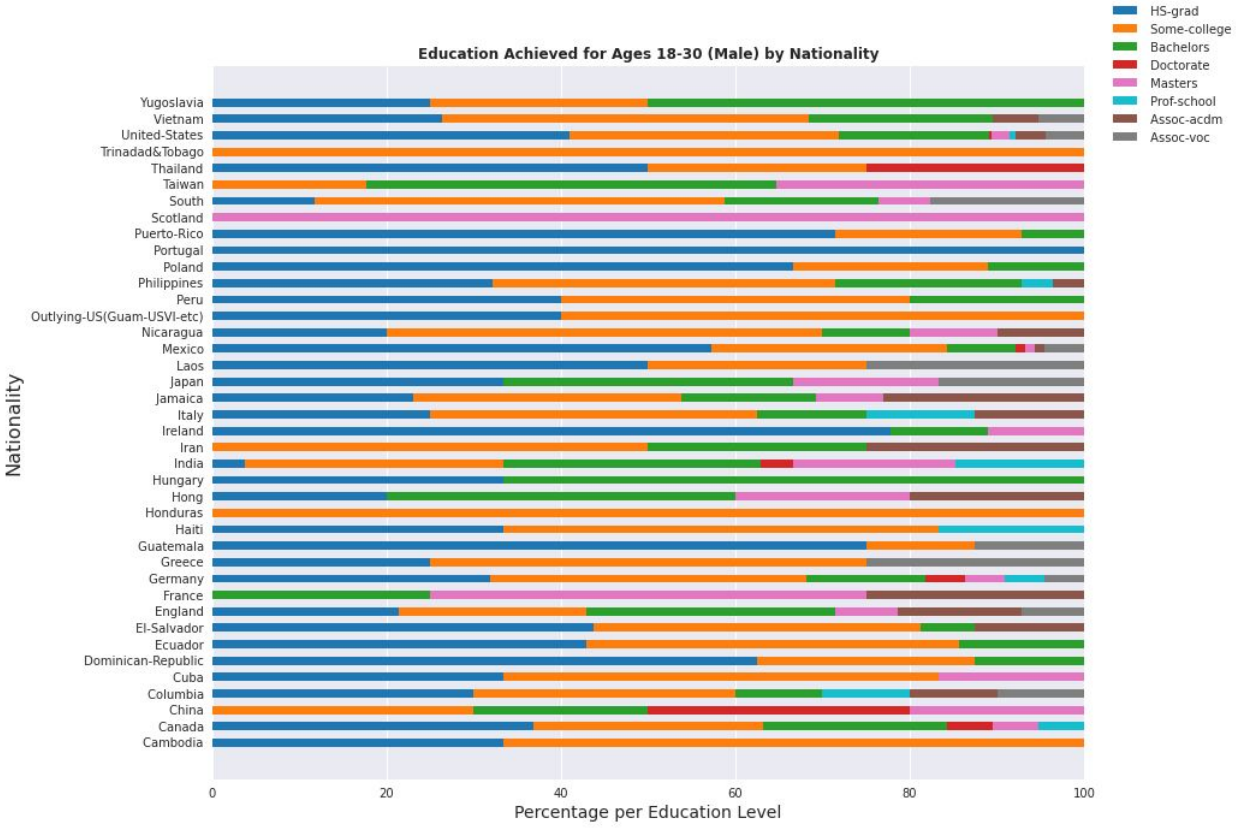
Appendix

I. Visualization#1

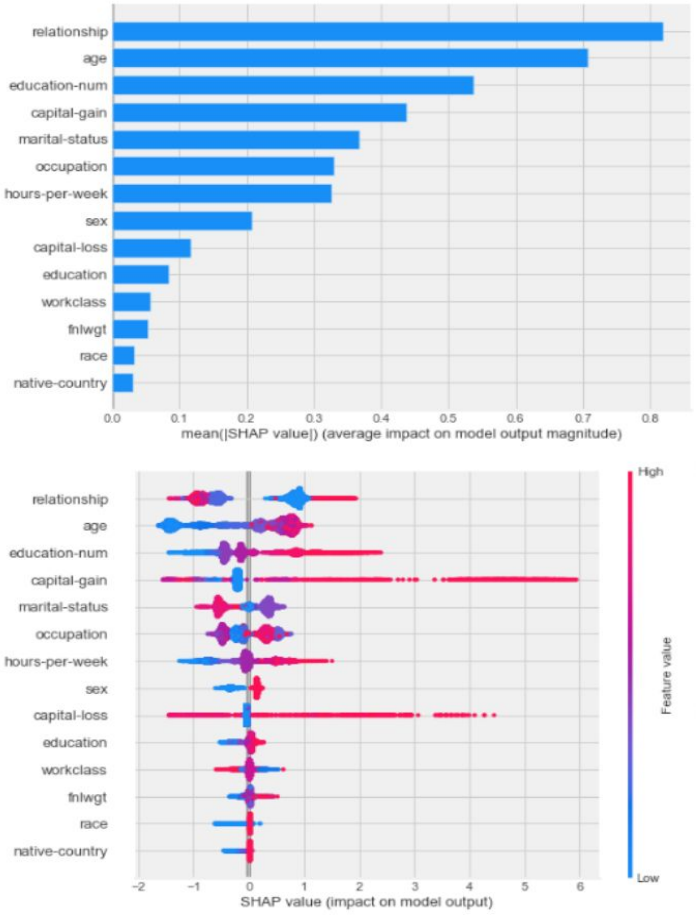





Education Achieved for Ages 18-30 (Female) by Nationality
Nationality
Yugoslavia
Vietnam
United-States
Thailand
Taiwan
South
Scotland
Puerto-Rico
Portugal
Poland
Philippines
Peru
Outlying-US(Guam-USVI-etc)
Nicaragua
Mexico
Laos
Japan
Jamaica
Italy
Ireland
Iran
India
Hungary
Hong
Honduras
Haiti
Guatemala
Greece
Germany
France
England
El-Salvador
Ecuador
Dominican-Republic
Cuba
Columbia
China
Canada
Cambodia
0
20
40
60
80
100
Percentage per Education Level

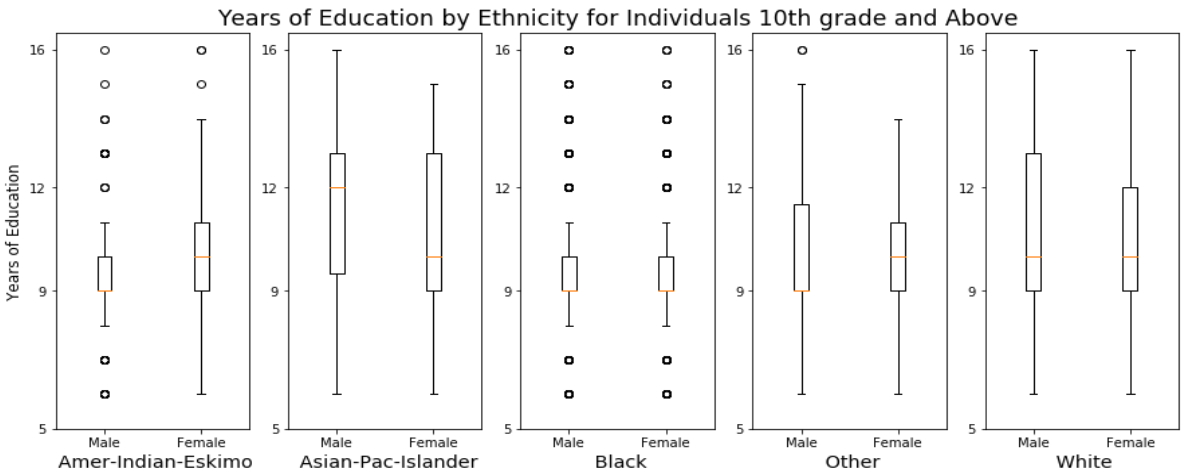

Education Achieved for Ages 18-30 (Male) by Nationality
Nationality
Yugoslavia
Vietnam
United-States
Trinadad&Tobago
Thailand
Taiwan
South
Scotland
Puerto-Rico
Portugal
Poland
Philippines
Peru
Outlying-US(Guam-USVI-etc)
Nicaragua
Mexico
Laos
Japan
Jamaica
Italy
Ireland
Iran
India
Hungary
Hong
Honduras
Haiti
Guatemala
Greece
Germany
France
England
El-Salvador
Ecuador
Dominican-Republic
Cuba
Columbia
China
Canada
Cambodia
0
20
40
60
80
100
Percentage per Education Level
HS-grad
Some-college
Bachelors
Doctorate
Masters
Prof-school
Assoc-acdm
Assoc-voc

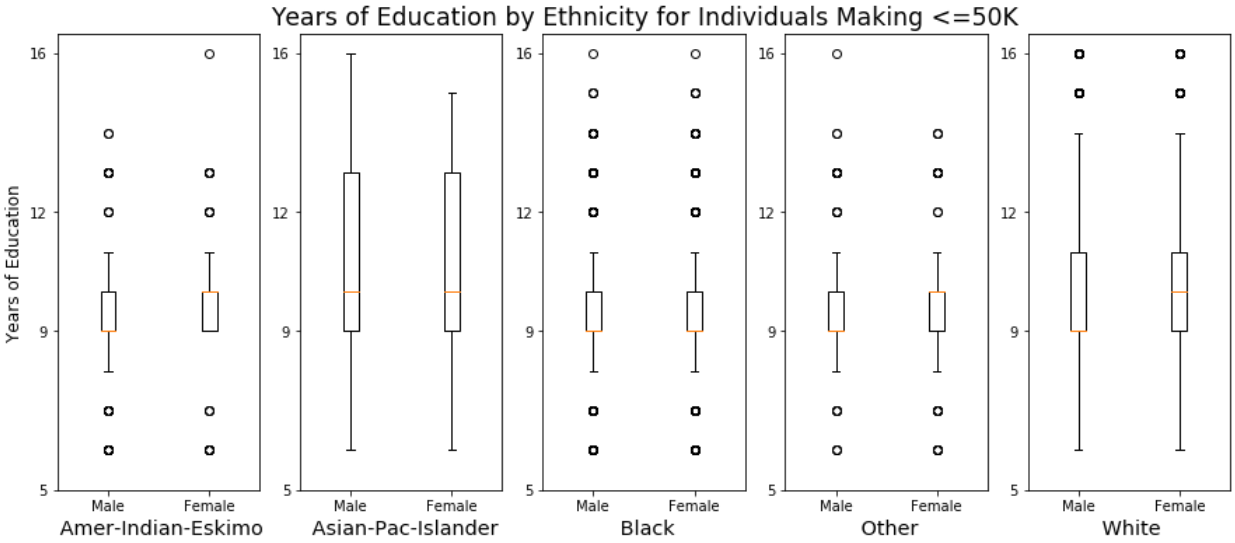
II. Visualization#2

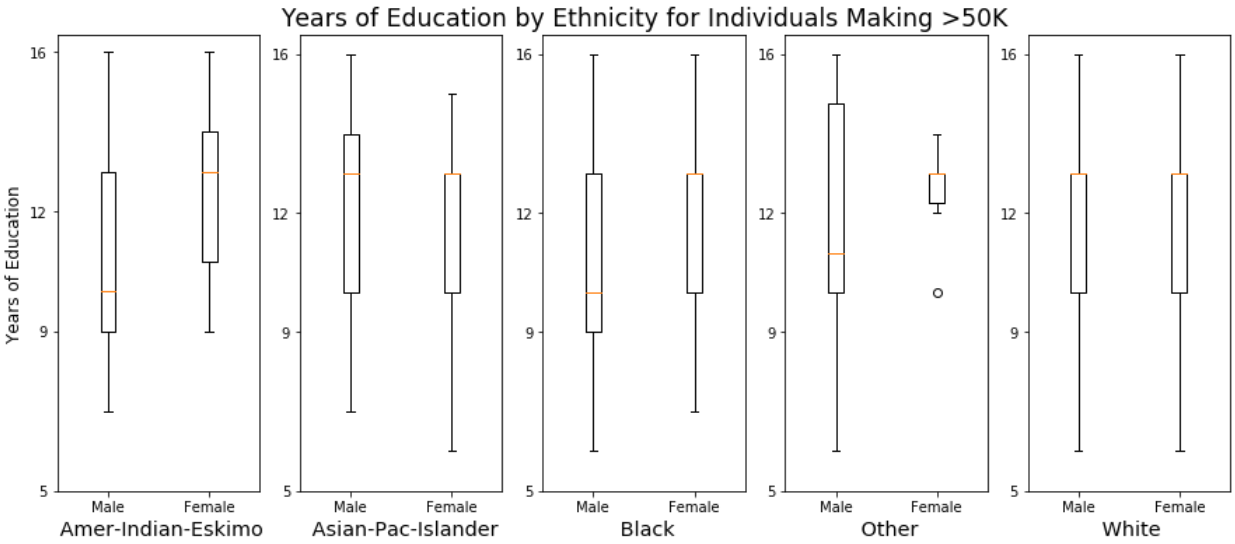


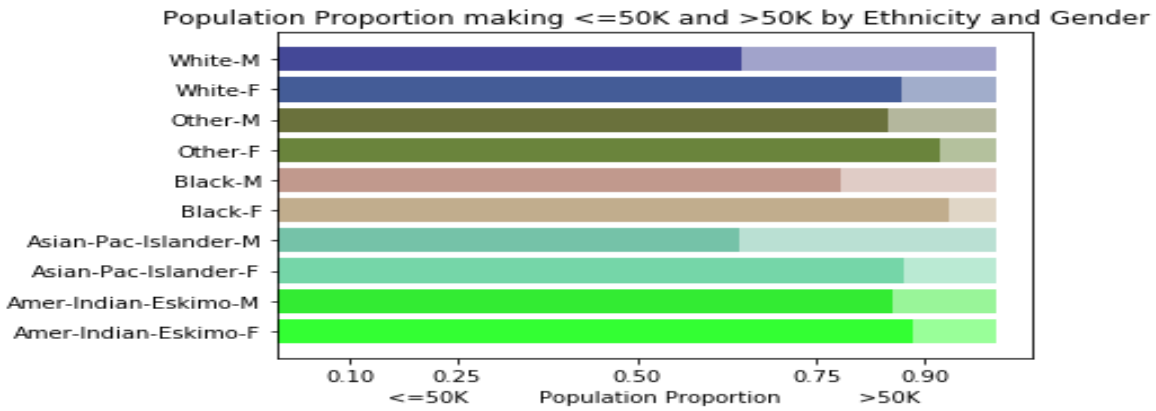
	column_name	shap_importance
7	relationship	0.746656
0	age	0.626939
4	education-num	0.515205
10	capital-gain	0.433327
5	marital-status	0.369107
6	occupation	0.340632
12	hours-per-week	0.339498
9	sex	0.200037
11	capital-loss	0.126448
3	education	0.0864223
2	fnlwgt	0.0674727
1	workclass	0.0562161
8	race	0.0274393
13	native-country	0.0158496

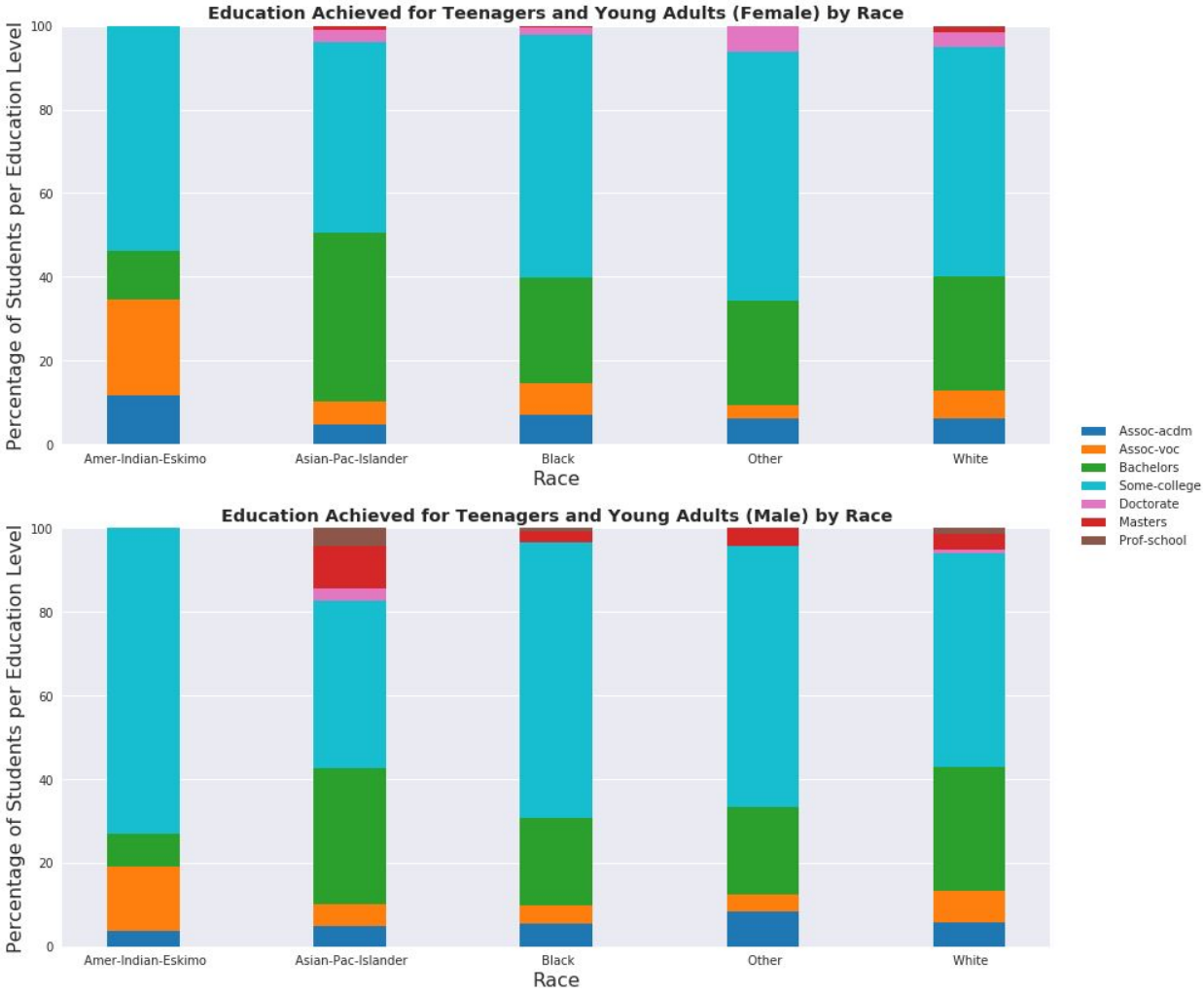
III. Visualization#3



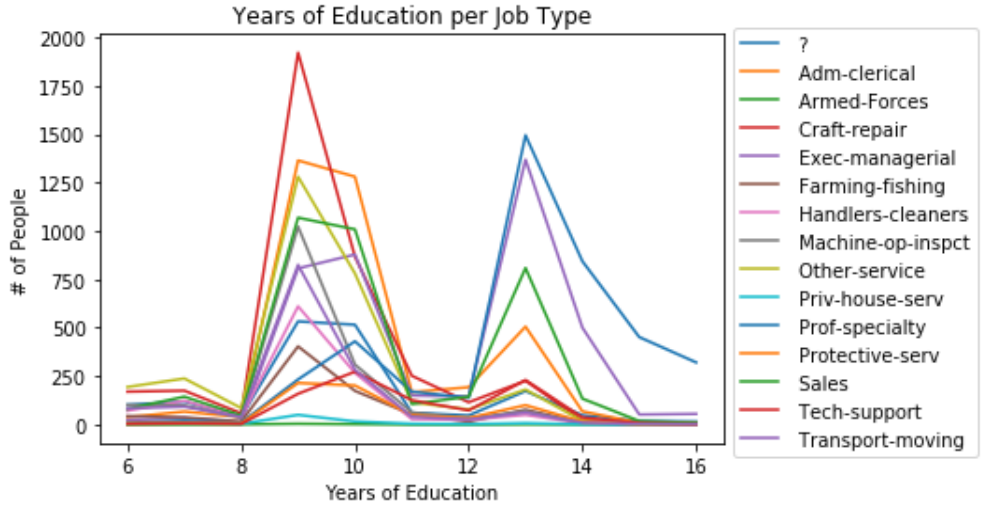

Years of Education by Ethnicity for Individuals Making <=50K
Years of Education
16
12
9
5
Male
Female
Amer-Indian-Eskimo
Asian-Pac-Islander
Black
Other
White

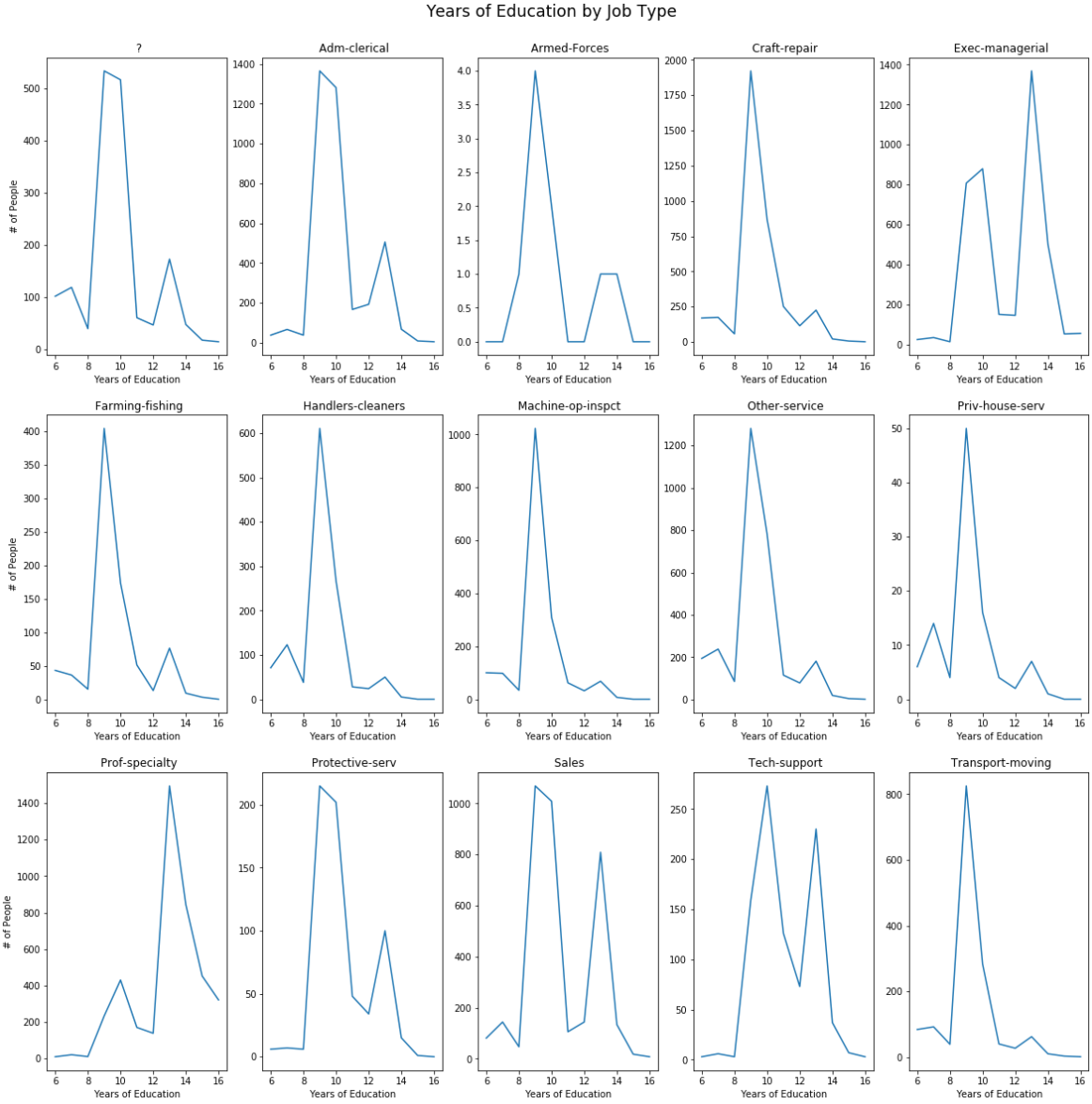

Years of Education by Ethnicity for Individuals Making >50K
Years of Education
16
12
9
5
Male
Female
Amer-Indian-Eskimo
Asian-Pac-Islander
Black
Other
White

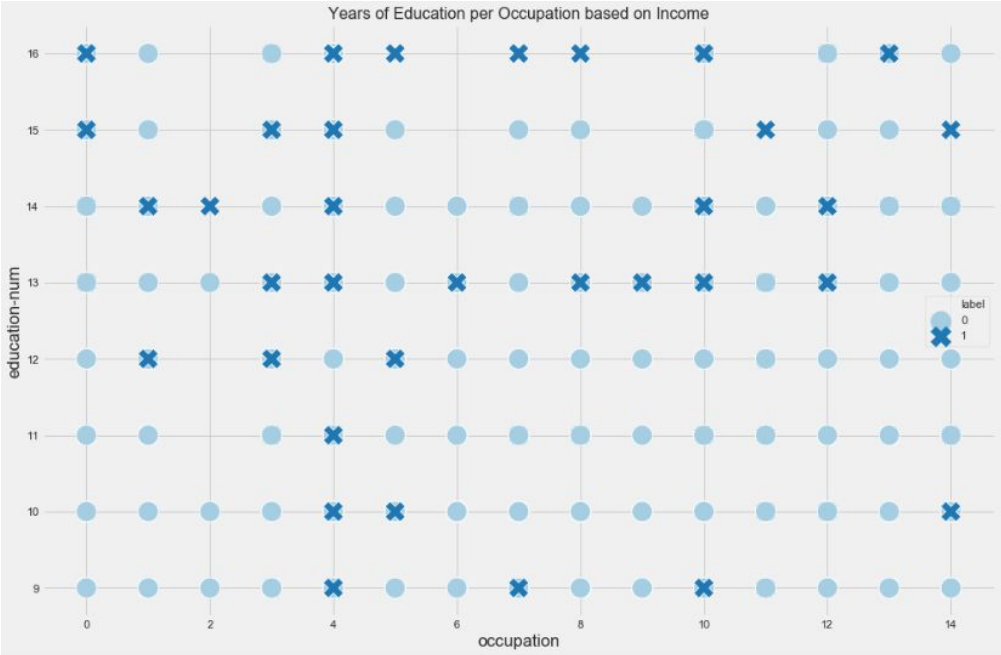

Population Proportion making <=50K and >50K by Ethnicity and Gender
White-M
White-F
Other-M
Other-F
Black-M
Black-F
Asian-Pac-Islander-M
Asian-Pac-Islander-F
Amer-Indian-Eskimo-M
Amer-Indian-Eskimo-F
0.10
0.25
0.50
0.75
0.90
<=50K
Population Proportion
>50K



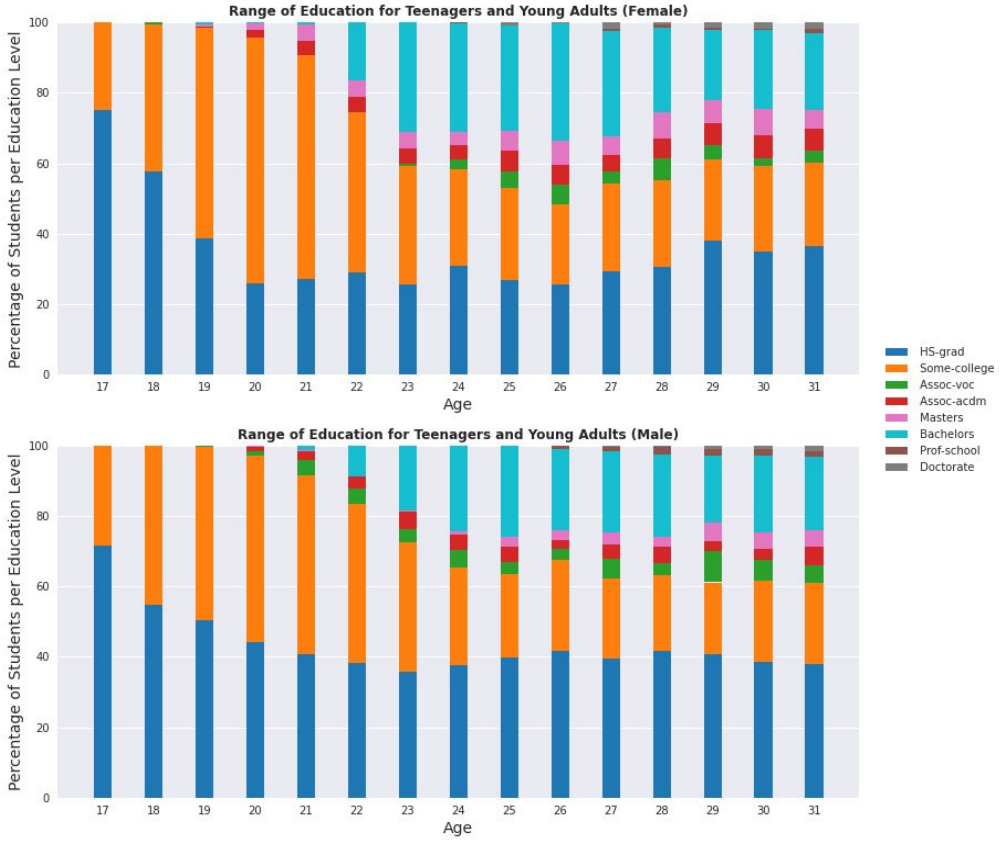
IV. Visualization#4

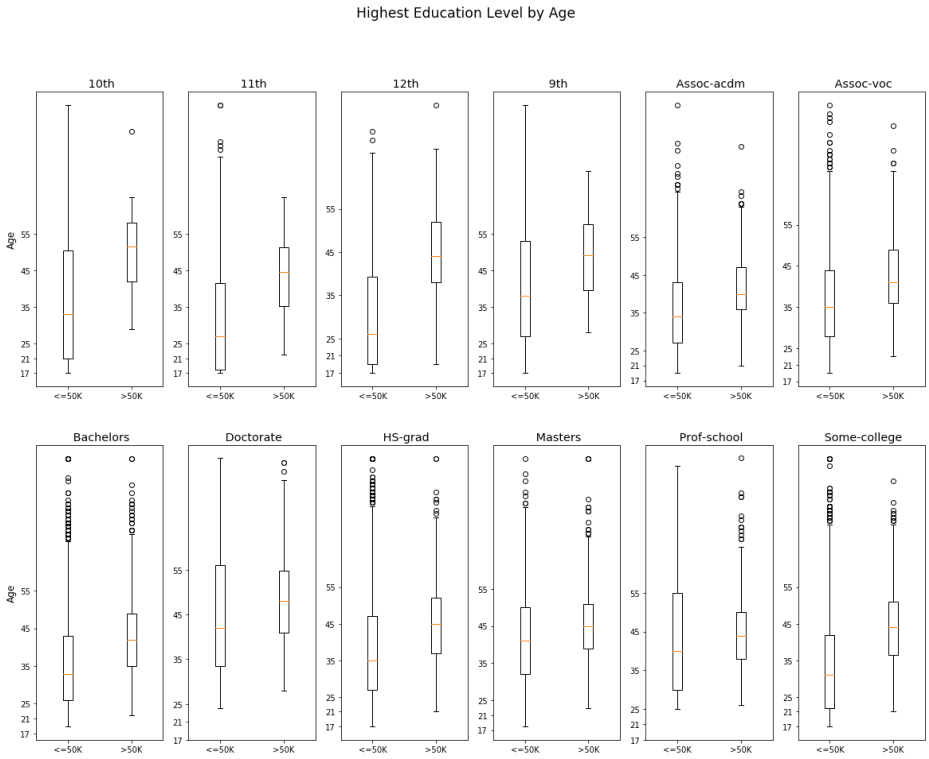


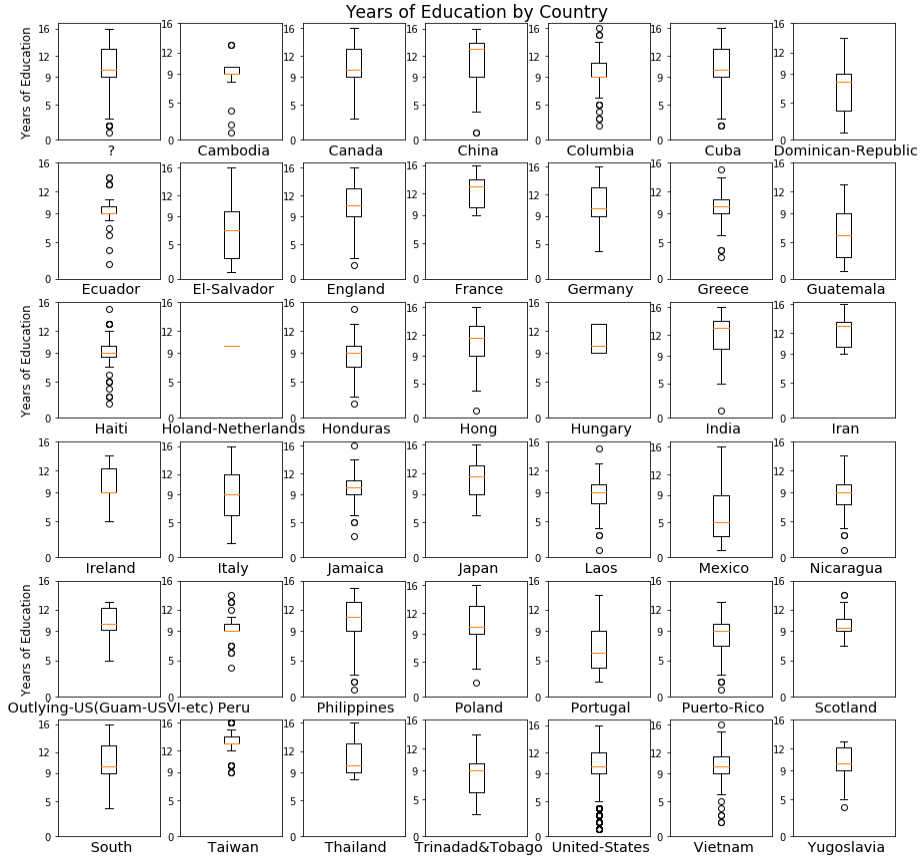

Years of Education by Job Type
?
Adm-clerical
Armed-Forces
Craft-repair
Exec-managerial
Farming-fishing
Handlers-cleaners
Machine-op-inspct
Other-service
Priv-house-serv
Prof-specialty
Protective-serv
Sales
Tech-support
Transport-moving
# of People
Years of Education



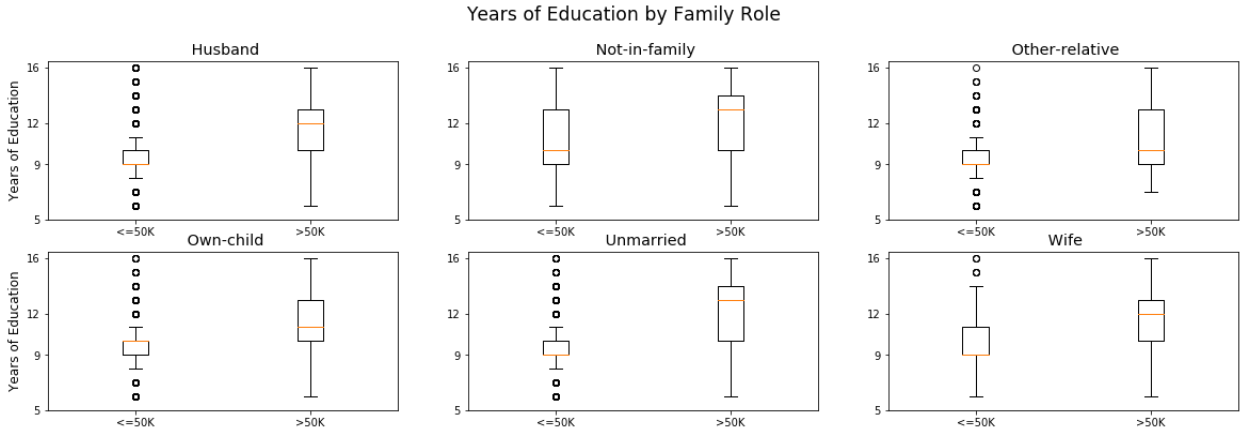
V. Visualization#5

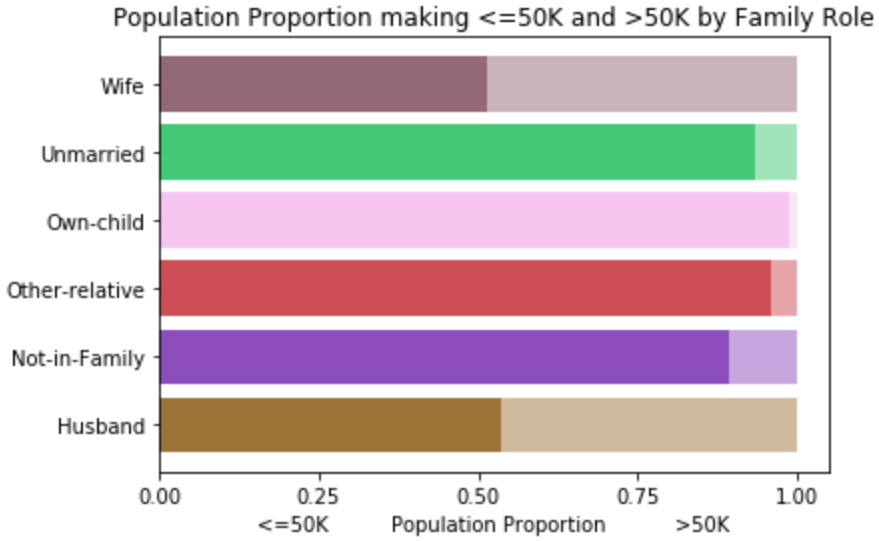



Highest Education Level by Age
10th
11th
12th
9th
Assoc-acdm
Assoc-voc
Bachelors
Doctorate
HS-grad
Masters
Prof-school
Some-college
Age
<=50K
>50K

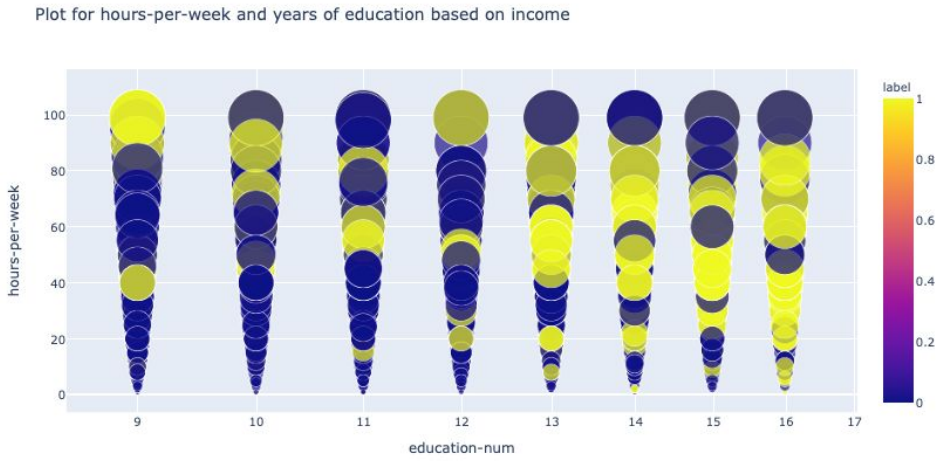

Years of Education by Country
Years of Education
?
Cambodia
Canada
China
Columbia
Cuba
Dominican-Republic
Ecuador
El-Salvador
England
France
Germany
Greece
Guatemala
Haiti
Holand-Netherlands
Honduras
Hong
Hungary
India
Iran
Ireland
Italy
Jamaica
Japan
Laos
Mexico
Nicaragua
Outlying-US(Guam-USVI-etc)
Peru
Philippines
Poland
Portugal
Puerto-Rico
Scotland
South
Taiwan
Thailand
Trinadad&Tobago
United-States
Vietnam
Yugoslavia

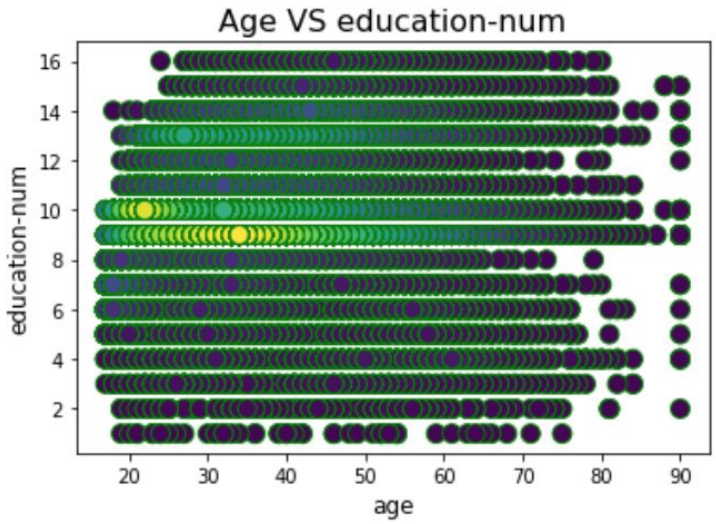
VI. Visualization#6

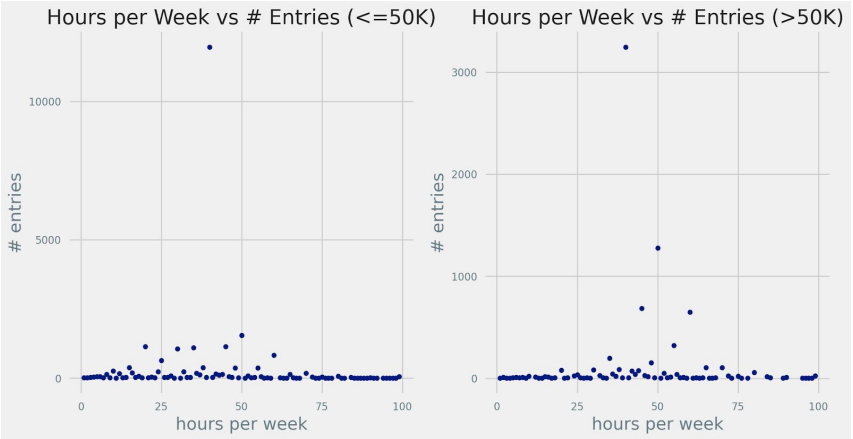


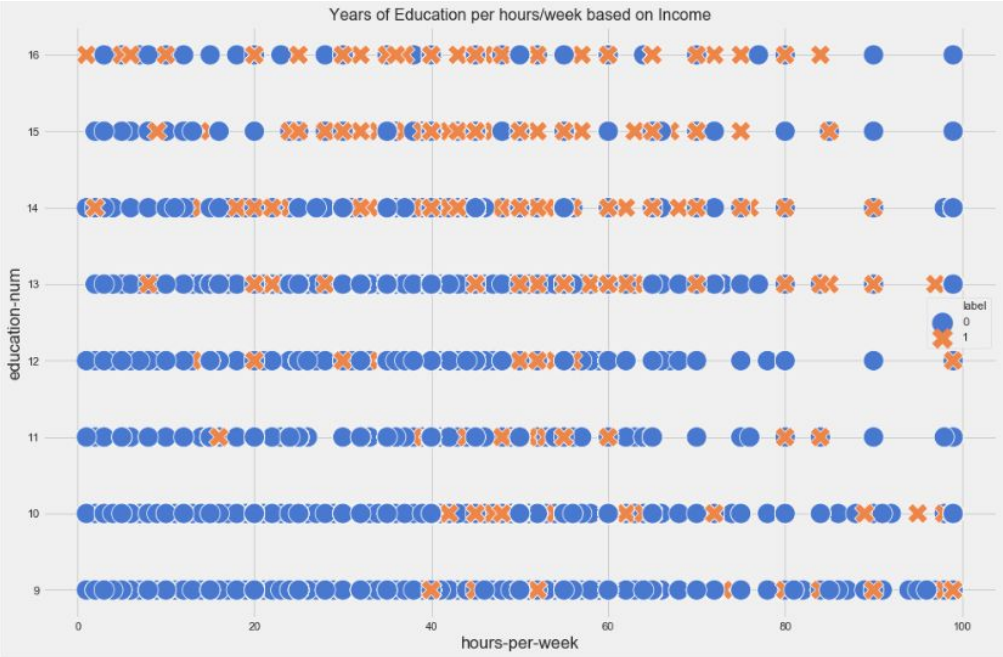


VII. Visualization#7

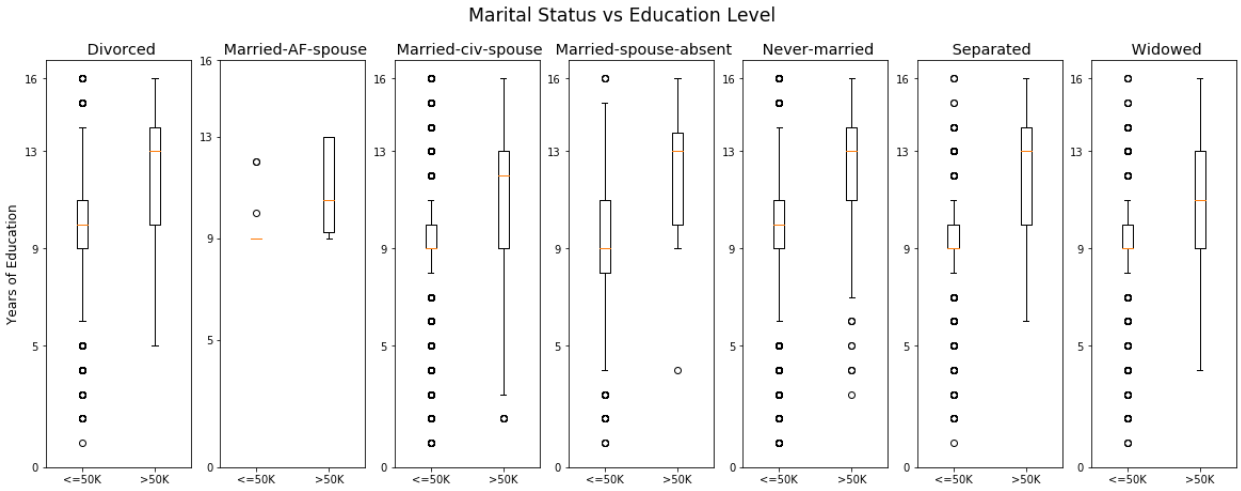


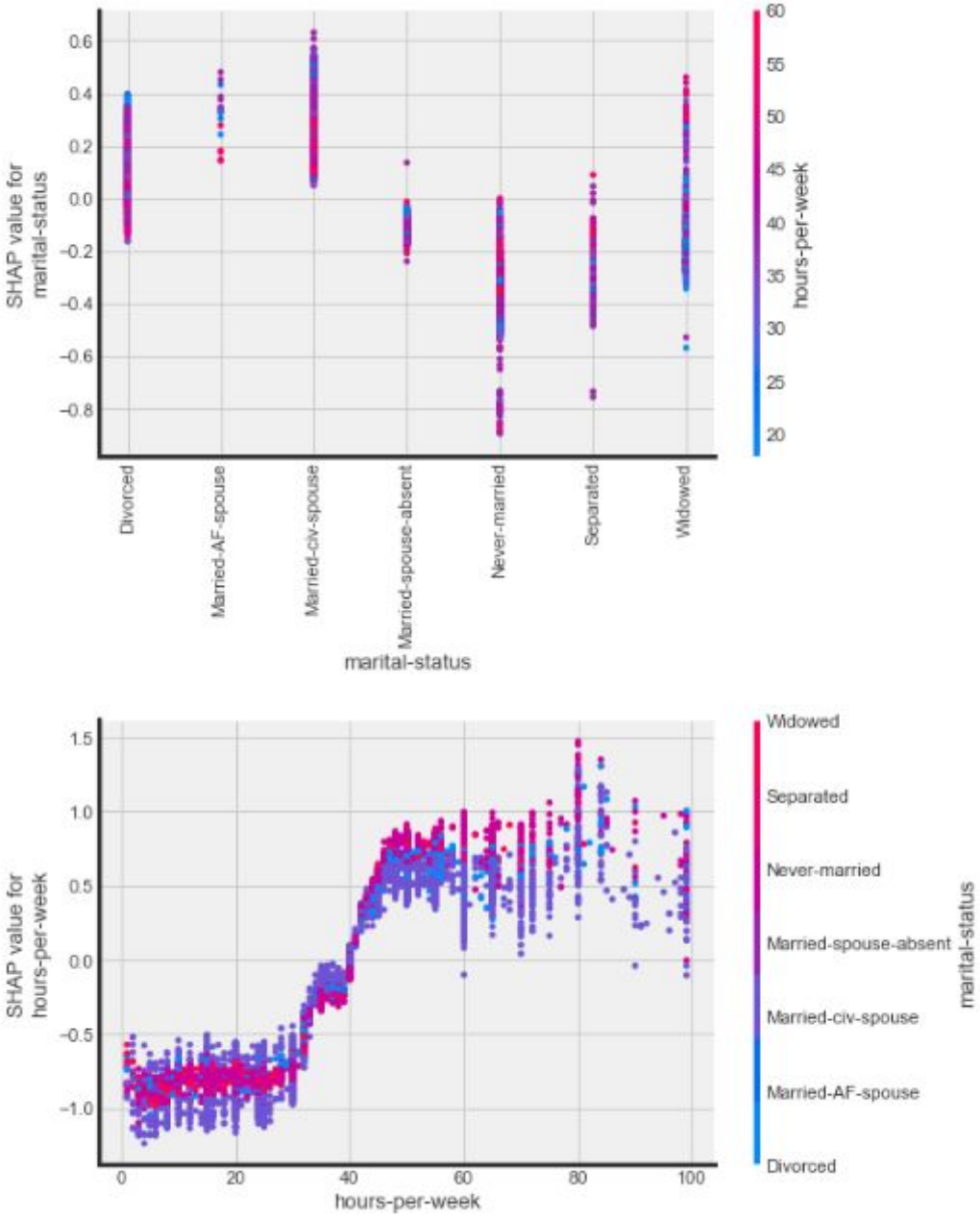






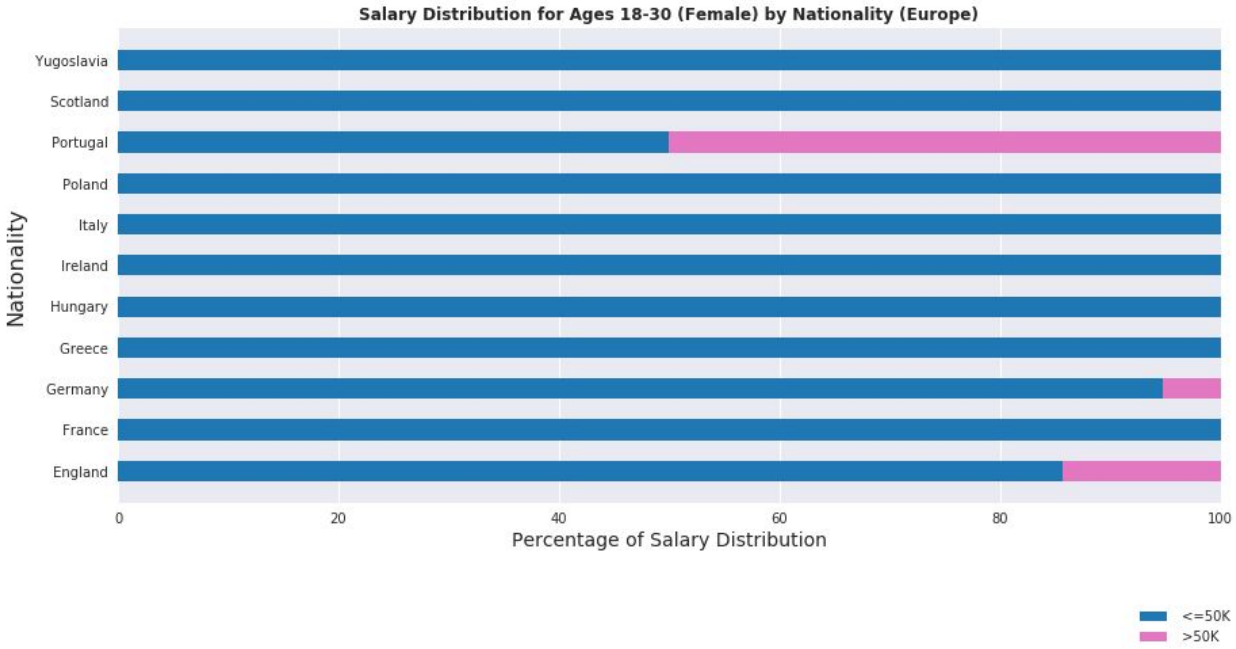
VIII. Visualization#8





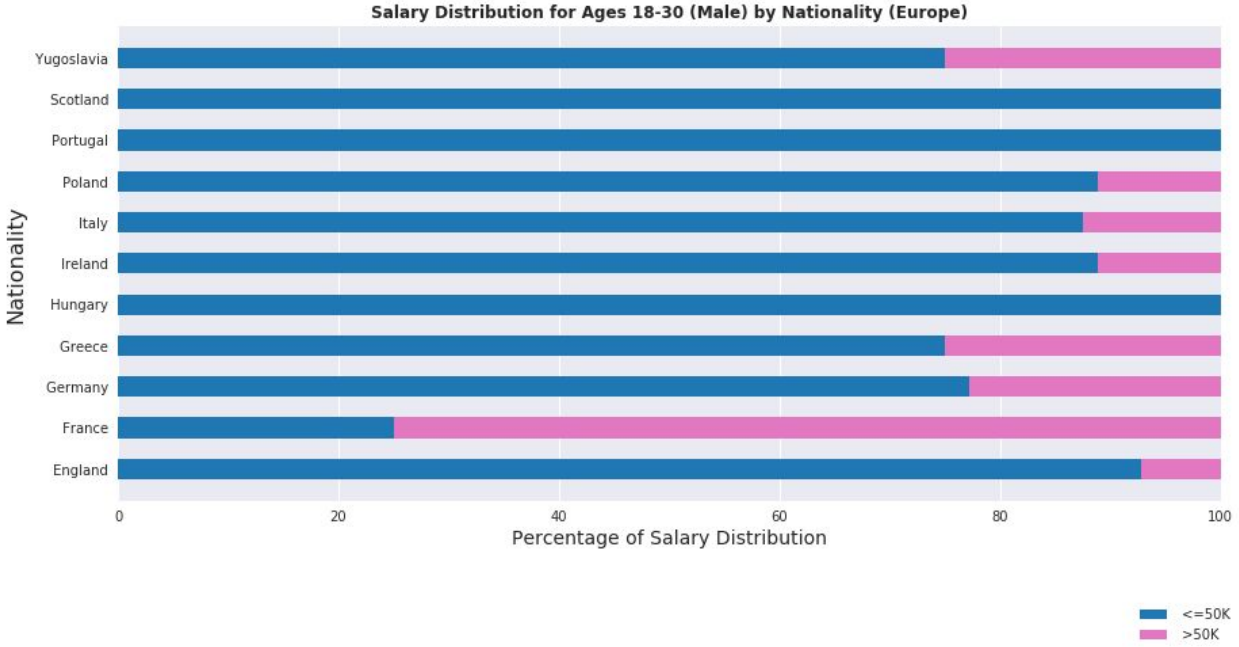
Higher SHAP value indicates Salary Distribution is on >50K.

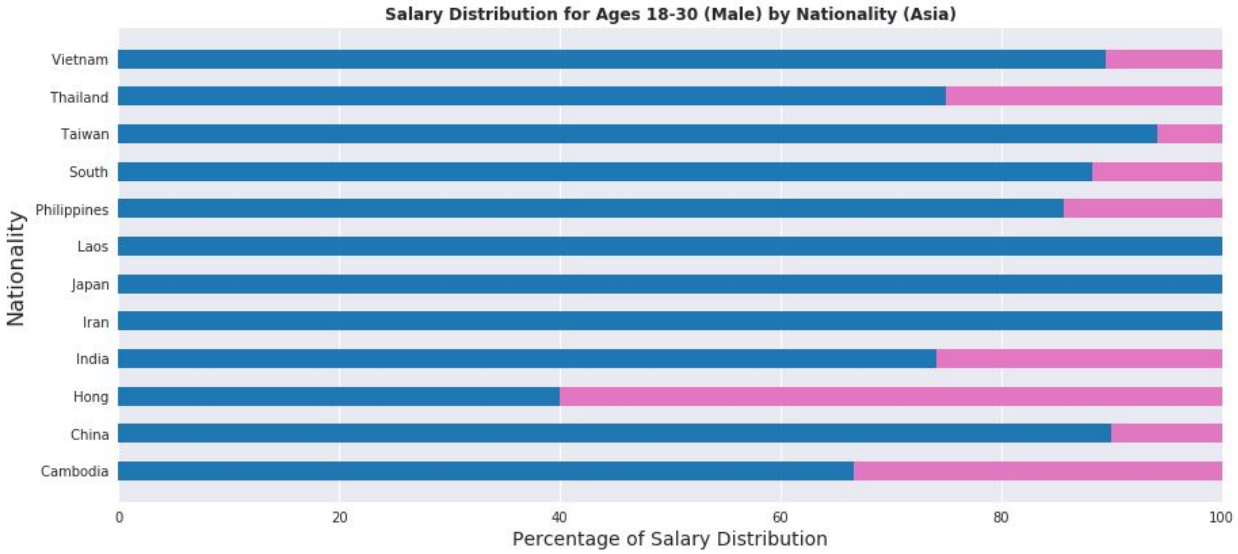
IX. Visualization#9

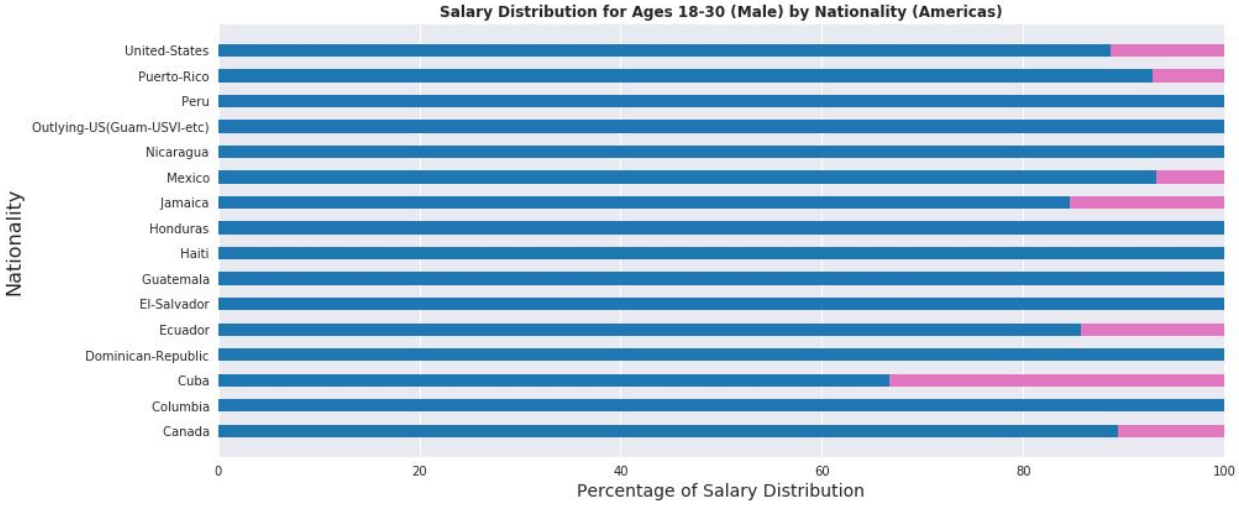

Salary Distribution for Ages 18-30 (Female) by Nationality (Europe)
Yugoslavia
Scotland
Portugal
Poland
Italy
Ireland
Hungary
Greece
Germany
France
England
Nationality
0
20
40
60
80
100
Percentage of Salary Distribution
<=50K
>50K


Salary Distribution for Ages 18-30 (Female) by Nationality (Asia)
Vietnam
Thailand
Taiwan
South
Philippines
Laos
Japan
Iran
India
Hong
China
Cambodia
Nationality
0
20
40
60
80
100
Percentage of Salary Distribution

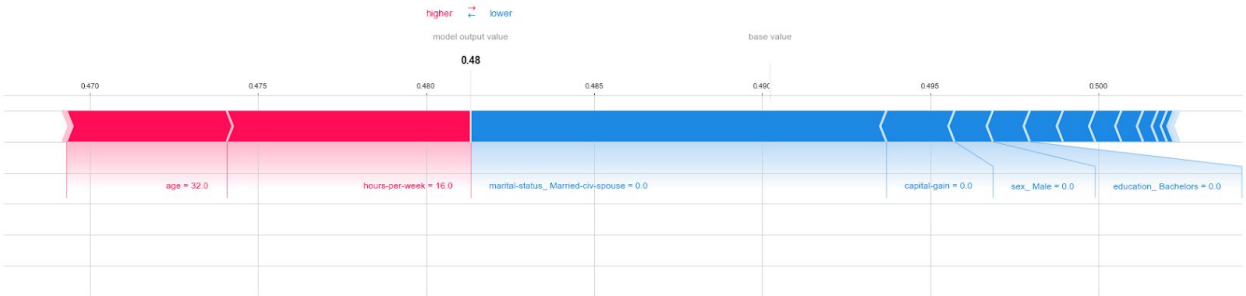

Salary Distribution for Ages 18-30 (Female) by Nationality (Americas)
United-States
Puerto-Rico
Peru
Outlying-US(Guam-USVI-etc)
Nicaragua
Mexico
Jamaica
Honduras
Haiti
Guatemala
El-Salvador
Ecuador
Dominican-Republic
Cuba
Columbia
Canada
Nationality
0
20
40
60
80
100
Percentage of Salary Distribution

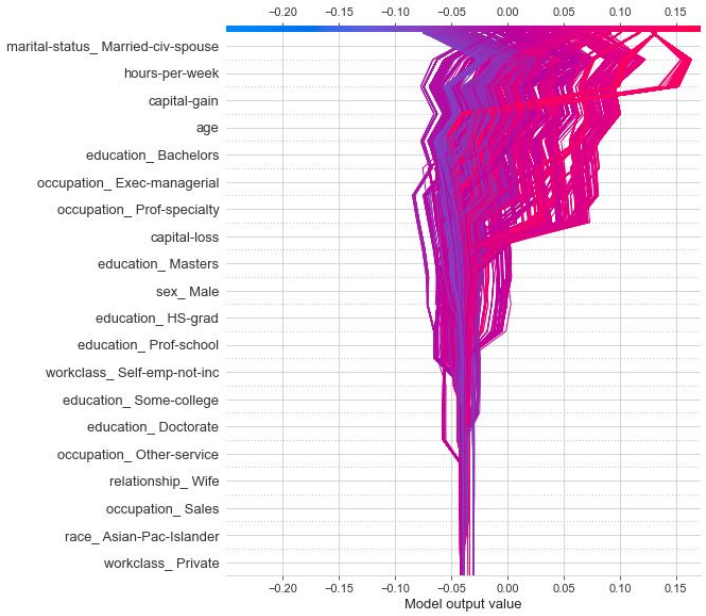

Salary Distribution for Ages 18-30 (Male) by Nationality (Europe)
Yugoslavia
Scotland
Portugal
Poland
Italy
Ireland
Hungary
Greece
Germany
France
England
Nationality
0
20
40
60
80
100
Percentage of Salary Distribution
<=50K
>50K





X. Visualization#10





A: prospective undergraduates

- ***marital status***: not married
- ***workhours/week*** : under 30
- ***capital gain/loss***: none
- ***age:*** under 25
- ***education:*** no Bachelor degree
- ***occupation:*** non-managerial, low/medium skill.
- ***gender:*** most probably woman
- ***income:*** <= $50.000/year

B: prospective graduates

- ***marital status***: married
- ***workhours/week*** : 40+
- ***capital gain/loss***: yes, notable
- ***age:*** 25+
- ***education:*** at least a BA degree
- ***occupation:*** managerial, or specialized/high skill.
- ***gender:*** most probably man
- ***income:*** > $50.000/year