CSE-578 DataVisualization Portfolio

Sailendri Gunnia Ravi kumar
sgunniar@asu.edu

***Abstract*—This portfolio includes the visualizations, which shows the parameters influencing salary. These parameters are taken from the United census data. With the parameters given by XYZ, UVW college marketing team predicts the salary of an individual. These results helps the marketing team to bolster the UVW college enrollment.**

Keywords—visualization, libraries, scenarios, inferences, attributes.

I. Introduction

XYZ corporation has signed a new project with UVW college. XYZ develops marketing profiles on people and sells them to different companies. UVW college is interested in enhancing their enrollments and so the marketing team is planning to create an application, which helps the team to forecast the salary of an individual. Salary is chosen as the key factor(50K). XYZ is expected to provide few vital parameters from "United States Census Bureau" data which helps the marketing team to obtain the salary values with higher degree of accuracy. These parameters are selected through interesting visualizations of patterns and relations from the data, which influence salary to a great extent. These parameters serve as input to the model being constructed by the marketing team. These results help the college to change the course structure, tuition fees according to different categories of people. Furthermore, the marketing team can reach out to potential individuals with higher enrollment probability.

II. EXPLANATION OF SOLUTION

A. Scenario 1

The Marketing team wants to know how Work-class, Occupation and Salary are related. The solution is divided into 2 parts. The first part of the story deals with relationship between the work class and salary. For this, a multivariate bar chart is selected. This gives us the population count on each category for different salary classes. At first stacked bar chart was selected for this scenario. But due to the drastic changes in values of private and other workclasses, the counts for each workclasses were not clear and the chart was not visibly appealing. The work classes 'Without-pay', 'Never-worked' and '?' are removed as they have very less data and they are not intuitive. The second part focuses on the salary distribution of the people belonging to different occupations of Private work class. This is because, the private workclass had more count with <=50K salary. For the visualization part, mosaic plot is used. It is an ideal plot for representing different classes of data. It also overcomes the disadvantage of pie chart by expressing more groups efficiently. For mosaic-plot, a new data frame with records corresponding to occupation and salary is constructed. The counts for both salary categories are appended. 'matplotlib.pyplot' and '*statsmodels.graphics*.mosaic' plot libraries[4] are used for constructing bar charts and mosaic plots respectively.

Fig 1.1

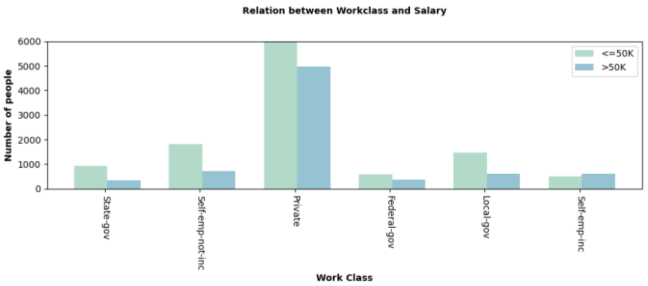
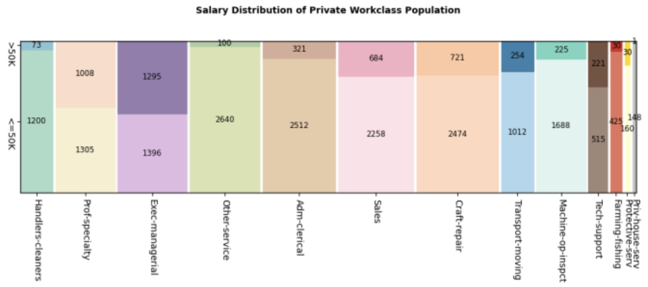


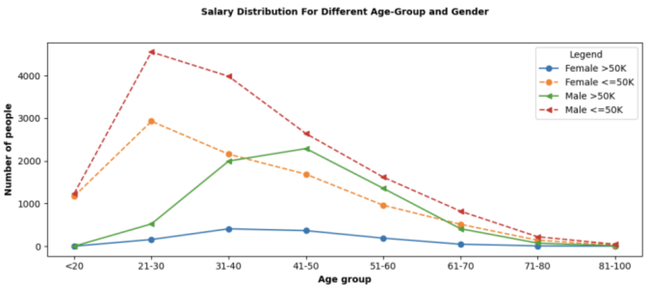
Fig 1.2



B. Scenario 2

The Marketing team is interested in visualizing the relationship between Age, Gender and Salary. This scenario deals with population count for different genders and different salary categories i.e.., >50K and <=50K. For this scenario line chart is used because, it gives a clear and distinct representation for all classes (Male <= 50K, Male >50K, Female <= 50K, Female >50K). The data is organized from the main table by aggregating the whole record with 'Sex' and 'Salary' attributes. Then for each combination of gender and salary, the data is categorized based on the age group. As data for age group is random, we are grouping them into several buckets ('<20','21-30','31-40','41-50','51-60','61-70','71-80','81-100'). 'matplotlib.pyplot' library is used.

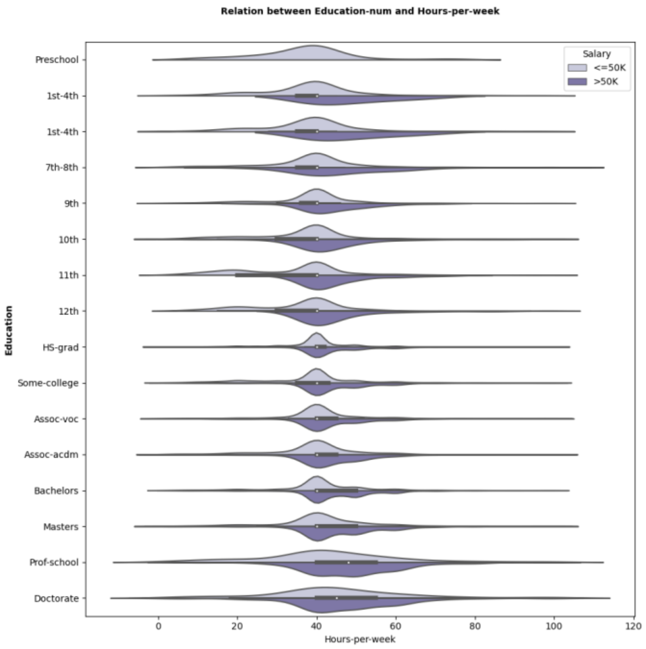
Fig 2



C. Scenario 3

The Marketing executives are curious in understanding the relationship between Education, Hours-per-Week and Salary. This visualization deals with the number of working hours of people with different level of education. For the visualization, we use violin plot, by taking hours per week on x-axis and different education levels on y-axis. Hue parameter is used to differentiate the salary categories. The original data frame is added along with hue value, color palette (‘Purples’) and split parameter as True. Split parameter (when given True) will show both salary categories on either side of violinplot. ‘seaborn’ library is used[3]. Education-num is used for plot construction and labels of y-axis are replaced with their corresponding education levels.

Fig 3



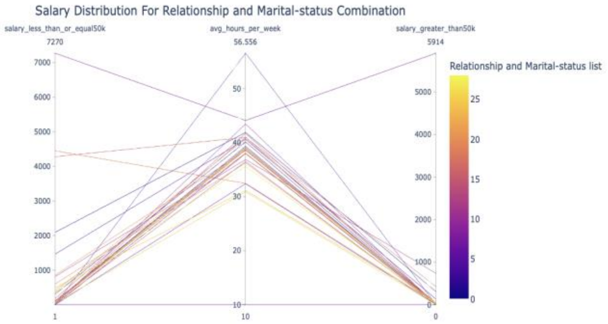
D. Scenario 4

The UVW college marketing team is interested in the association of Marital status, Relationship and Salary. This visualization deals with the different combinations of ‘Marital status’ and ‘Relationship’ and their relationship with the two salary categories and average hours per week.

For this scenario, parallel co-ordinate plot is used because, it is a powerful plot for representing multivariate data. It gives a separate connected line segment for all combinations, and the patterns and similarities between different combinations can be found. The selected parameters can be of any range, and it can be normalized. The original data frame is grouped by ‘Marital status’, ‘Relationship’ and ‘Salary’. For each combination, we group corresponding rows and note counts for the >50K and <=50K salary categories. The average hours-per-week is calculated separately for all the combinations of ‘Marital status’ and ‘Relationship’. ‘px.parallel_coordinates()’ function is used to construct plot[2].

Names of the combination are encoded from 0 to 28 numbers.{ 'Divorced & Not-in-family':0, 'Divorced & Other-relative':1, 'Divorced & Own-child':2, 'Divorced & Unmarried':3, 'Married-AF-spouse & Husband':4, 'Married-AF-spouse & Otherrelative':5, ’Married-AF-spouse & Own-child':6, 'Married-AF-spouse & Wife':7, 'Married-civspouse & Husband':8, 'Married-civ-spouse & Not-in-family':9, 'Married-civ-spouse & Otherrelative': 10, 'Married-civ-spouse & Ownchild':11, 'Married-civ-spouse & Wife':12, 'Married-spouse-absent & Not-in-family':13, 'Married-spouse-absent & Other-relative':14, 'Married-spouse-absent & Own-child':15, 'Married-spouse-absent & Unmarried':16, 'Never-married & Not-in-family':17, 'Nevermarried & Other-relative':18, 'Never-married & Own-child':19, 'Never-married & Unmarried':20, 'Separated & Not-in-family':21, 'Separated & Other-relative':22, 'Separated & Own-child':23, 'Separated & Unmarried':24, 'Widowed & Not in- family':25, 'Widowed & Other-relative':26,'Widowed & Own- child':27, 'Widowed & Unmarried':28}.

Fig 4



E. Scenario 5

The UVW team wants to know the link between Country, Race, and Income. This visualization deals with the population count of ‘Asian-Pac-Islander’ race in different countries. The people of this race are spread across the different countries. Choropleth map is used for this visualization. Because this map best explains the data across different geographic regions. ‘plotly.graph_objects’ library is used. To use it in offline we are making init_notebook_mode(connected=True)[1] The geometry of the polygons (countries) is taken from the ‘naturalearth_lowres’ dataset which is in-built in geopandas library. Our original dataframe, is grouped by ‘Race’ and ‘Salary’. The population count for each combination is noted for >50K and <=50K categories. ‘Natural Earth projection’ is used. Using go.Figure(), we combine the data and layout dictionary. With iplot(choromap), we plot two visualizations. One for salary<=50K and other for salary >50K.

Fig 5.1

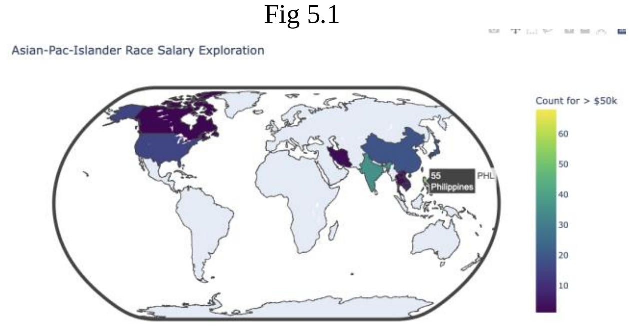
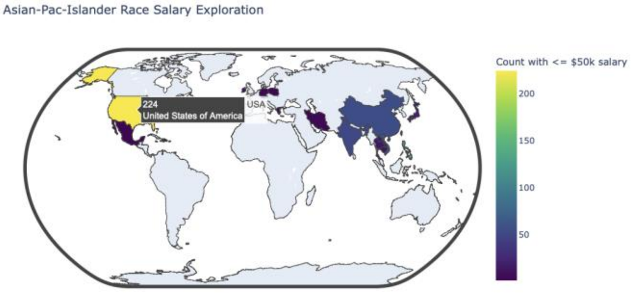


Fig 5.2


Asian-Pac-Islander Race Salary Exploration
224
United States of America
USA
Count with <= $50k salary
200
150
100
50

III. DESCRIPTION OF RESULTS

A. Scenario 1

We can observe that individuals belonging to professions such as "Prof-Specialty" and "Exec-managerial" are expected to get higher income when compared to the population in other occupations. These people have a higher probability of transitioning from the <=50K salary to >50K salary category. So, the college can plan special courses for the people of Private work-class and the above-mentioned occupations. Because people might always prefer to upgrade themselves instead of changing their entire stream of occupation for getting higher salary. The special courses can be of types nano-degree program, certification, hands-on-training etc.

B. Scenario 2

From the visualization for scenario 2, we can infer that, male population in the age group of <=50 earn more compared to female population. Furthermore, the count of male population belonging to the salary category >50K and age group of 41-50 is greater and the count of female population belonging to the salary category >50K and age group of 31-40 is greater. From this we can understand that, women tend to earn higher salaries earlier in their life. Women population (<1000) total count can be lesser than male population (2500). This can be used by the college to motivate female population and there is a higher scope that proper curriculum for female population of 31-40 age can aid them to get jobs with higher pay. In addition to that, there are more count for female of age 21-30 earning <=50K. The college can plan accordingly to train women of 21-30, as there is a higher probability for them to join >50K salary category in future.

C. Scenario 3

From the violin plot, we can infer that people with the education level of "Prof-School" work for an average of 50 hours per week, as shown by the median. Whereas, people with the level of "Doctorate" work for only 40 hours per week. From this, we can observe that, when a person upgrades from "Prof-school" to "Doctorate" can earn more and also reduce the working time up to 10 hours per week. We can also observe that, higher population are working around 40 hours per week. So, college can plan providing the courses which helps people to manage both their job and courses parallelly, for different categories like Masters, Prof-school and it can be like online courses or giving adequate time to complete the courses at their own pace. So, as a result people can work for the same 40 hours per week but for higher qualification jobs. This can greatly improve the salary of a person transitioning him/her with same amount of working hours.

D. Scenario 4

From the plot we can observe that the combination number 8 – "Married-civ-spouse and Husband" are earning more than all other combinations of Marital Status and Relationships. In addition to that there exists a positive correlation between Hours-per-week and Salary>50K.

E. Scenario 5

From the two choropleth maps for <=50K salary and >50K salary, we can hover on the areas of interest to check out the count of people belonging to "Asian-Pac-Islander" race with respective salary category. We can infer that, more people with <=50K salary are from United-States. We can also note that, individuals from Philippines earn >50K and among all the countries, their count is the highest. And another interesting fact is that Philippines has the second highest count of people earning <=50K too. From this, we can come to a conclusion that people of Philippines have a higher probability to transit from <=50K to >50K faster despite a sizable portion of them earning a lower salary.

IV. LESSONS LEARNED

- By the end of this individual project, I got a good hands-on experience on jupyter notebook, by importing and practicing on more libraries.
- It is understood that, selecting appropriate visualizations for different scenario, is heavily dependent on the underlying data.
- Pre-processing of data greatly influences the visualizations. Displaying data by removing nan data, unknown data or data with fewer records in some cases help us to gain more insights.
- Histogram gives a better idea on distribution of data with respect to specific attributes.
- We need to present visualizations with legends scales and proper title. We need to keep in mind Bertin's visual variables and select color schemes like rainbow, sequential, divergent for different data types like nominal, ordinal, interval and ratio.
- Hovering over visualizations to get info on places with different colors looks more attractive. As a result we can have more info on one visualization. This can be done with iplots like choromap, Sankey diagrams.
- I got to know how multivariate visualizations can be done, more effectively. I wish to implement it on different datasets and learn more on visualizations, as it helps us to take right decisions swiftly and with lesser efforts.

V. REFERENCES

[1] https://www.udemy.com/course/python-for-data-science-and-machine-learning-bootcamp/learn/lecture/5733392#overview

[2] https://plotly.com/python/I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[3] https://www.udemy.com/course/python-for-data-science-and-machine-learning-bootcamp/learn/lecture/5733294#overview

[4] https://www.coursera.org/learn/cse578/home/week/4