UVW College degree program marketing analysis

CSE 578 (Spring 2024) Final Project Report

Harshita Asthana
Arizona State University
hasthana@asu.edu

Data driven marketing is used by universities to bolster their degree program enrolments. XYZ corporation is helping the UVW college with data driven marketing analysis. We as an analyst working for XYZ are assigned to come up with the analysis that can help UVW college to make informed degree enrollment marketing decisions. No new dataset is gathered, and we are making use of an existing dataset from the United States Census Bureau. This dataset is used to establish the relationships and to come up with user stories for data visualizations that can communicate the relationships in easy-to-understand format. Data visualization can share the complex patterns for huge datasets which are otherwise hard to convey by looking at the whole dataset.

I. Introduction

As analysts working for XYZ corporation which is helping UVW college to make better data driven marketing decisions to bolster their degree program enrolments are assigned to come up with data insights. For this task UVW college has already identified a key attribute "salary" which must be included in our data analysis on the dataset that is supplied by the United States Census Bureau. No new dataset has been gathered and used in the analysis. This analysis final report is structured in a typical data science workflow and each step in the workflow is covered in upcoming sections.

II. Data Science Workflow

A. Problem Identification

UVW college has already identified the dataset from United States Census Bureau and a key attribute "salary" in the dataset which we have to use as an analyst working for XYZ corporation to come up with analysis which can be helpful for bolstering the degree program enrolments. Since it's an open-ended problem, we started the analysis by identifying the attributes which can be used to build good user stories. Based upon which we came up with the following user stories.

1) *How are salaries varying with **age**?*
2) *Salary variance by the **occupation.***
3) *How salaries are varying by **gender**.*
4) *How salaries are varying by **workclass.***
5) *How salaries are varying for the individuals having **native_country** not as USA.*
6) *How **education** and **education_num** linked with salaries. Does higher education guarantee more salaries?*
7) *Furthermore, can **workclass** be used along with **occupation** to determine the salary.*
8) *How **relationship** and **marital_status** can tell a lot about an individual's salary.*

Above mentioned stories are covering 9 attributes out of 14 total attributes to determine the salary of an individual. In some stories we either used single (univariate) or multiple (multi-variate) attributes to come up with salary analysis.

B. Data Acquisition

In this data analysis we are not acquiring any new data, instead we are relying on the dataset already identified by the UVW college and supplied by the United States Census Bureau. This dataset is publicly available from the UCI website as an *adult.data* file download. Along with the dataset UCI website also provides us with information about the various columns which are available in the dataset. For some columns like native_country we also have information of all the possible values that column can take.

C. Data Wrangling

We chose the Python environment for our analysis as it provides a lot of powerful libraries like numpy, pandas, matplotlib, scipy, etc. which makes analysis a lot easier. The dataset is first brought into system memory by making use of the *read_csv* API from the pandas library. Pandas API reads the dataset into an in-memory data frame from which we identified that there are 32561 rows and 15 columns. Out of these 15 columns we are excluding the *fnlwgt* from our analysis along with *capital_gain* and *capital_loss*. We also identified that there is a lot of noise in the dataset in the form of "*?*" values. These values need to be cleaned before doing any sort of analysis, so we replaced all such occurrences with Python None. There were no column headers in the original dataset due to which it was hard to deal with data based upon indexes in our code, so we added the column header to the pandas data frame.

D. Data Analysis and Modelling

From the cleansed dataset mentioned in the previous step we started our analysis on top of it. One of the initial steps was to group the dataset by salary and then use it with all other columns one by one and with a combination of columns. During this process we build multiple smaller data frames which are also used to build visualizations using matplotlib python library. Depending upon the columns used in some cases we built the univariate graphs like **line, bar, histogram**, etc. while in others built multivariate graphs like **parallel coordinate, scatter plot, mosaic plot**, **choropleth** etc.

E. Reporting And/Or Deployment

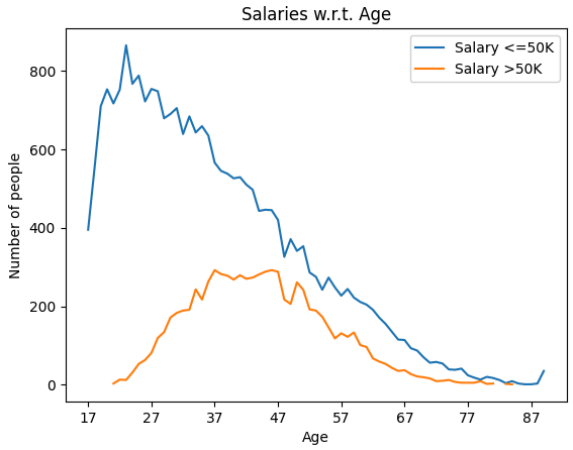
From the progress report to now we made significant progress in terms of resolving all open questions, coming up with all visualization and finally documenting them in the report. We also explored and used alternative visualization libraries like Bokeh, Seaborn, etc. for some of the user stories. Diagrams and the complete working code is hosted on Github and we have provided the links at the end of this doc so that we can look at high-definition visualization insights if needed be.

III. User Stories

In this section we will cover the user stories for which we were able to identify the insights by building appropriate visualizations. All of these insights are in relation to the key column "salary" which UVW college has already identified and on top of the data which was already pre-processed as mentioned in the Data Wrangling in the earlier section.

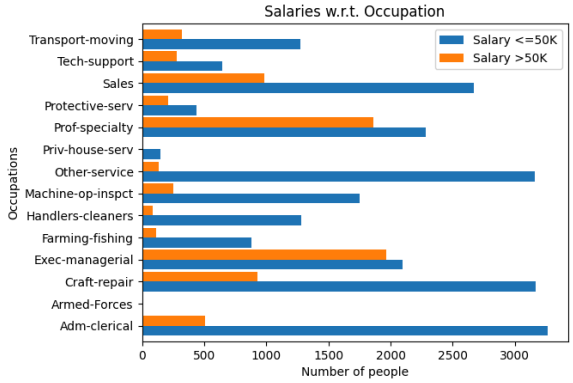
*1. How are salaries varying with **age**?*

Age, which is an integer data type in the dataset, is used to get insights on how salaries are varying by various age groups. For coming up with this insight we grouped the complete preprocessed dataset first by the income and later by the age using pandas dataframe. For all these groups we also identified the group size which we used in the visualization below. We went for a line graph because age is continuous and we can easily get the insight how salaries vary by age. Using this insight UVW college can reach individuals whose age is between 17 and 40, as line dips gradually after 40 years.



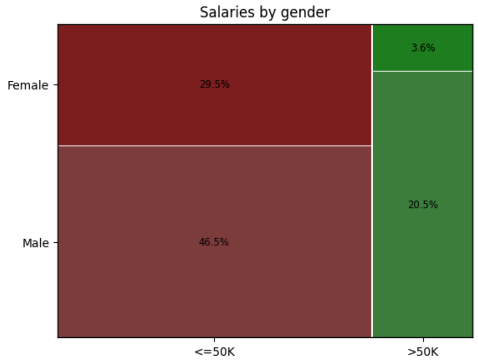
2. *Salary variance by the **occupation***

Similar to the previous user story, here also we grouped the complete preprocessed dataset first by income and later by the occupation using the pandas dataframe. For the grouped data we identified the group size and plotted in a horizontal bar chart so that we can figure out occupations with competitive salaries. Since occupation is not a continuous number like age we went for a bar chart. This occupation insight can be used by the UVW college to target the degree programs that will prepare the student to eventually secure a job in those occupations. For some occupations we don't have many data points to confidently deduce that those occupations will have enough jobs once candidates finish their degree program e.g. Priv-house-serv.



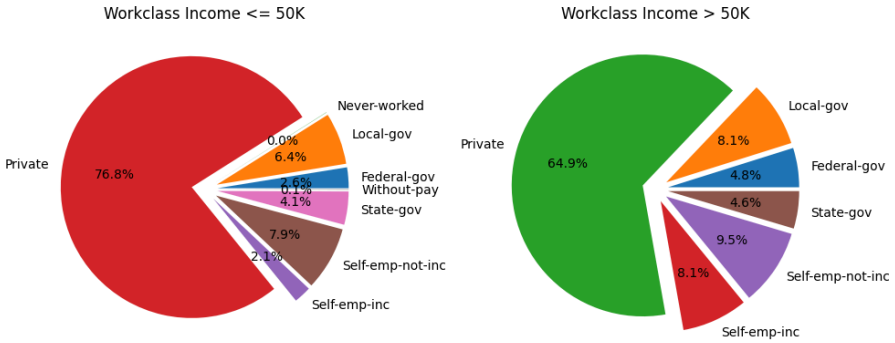
3. *Salary variance by the **gender***

For generating this insight we again used the complete preprocessed dataset which we first grouped by salary and later by gender. Once we had the grouped data, we plotted the group size as a mosaic plot from the *statsmodel* library. We used the mosaic plot here because our data in this case is mainly income and gender which are categorical in nature and both having only two possible values. This plot gives us insight into the biases which can come from the skewed data e.g. in the given dataset only 33.1% data points are for the female while remaining are for the male which is far from the real-world. Other than data only 3.6% of females are earning more than 50K, so UVW college can market the appropriate programs to the females that can help them secure more than 50K salary jobs.



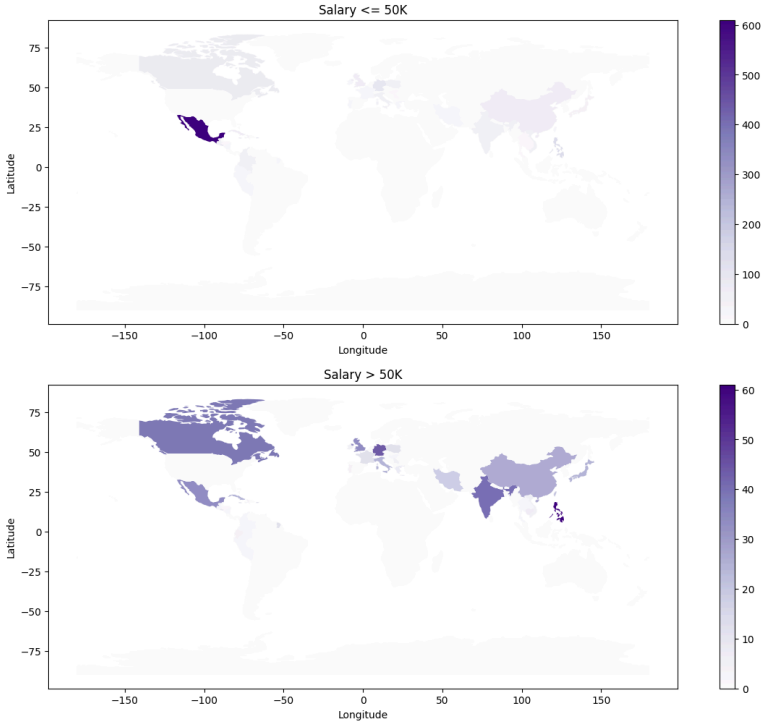
4. *Salary variance by the **workclass***

For this next user story for workclass we did the grouping of the complete preprocessed dataset first by the salary and later by the workclass. Once we had these groups, we plotted the pie chart as there were close to 6 and 8 workclass for each of the salary groups. Usually 7 categories are considered an ideal number of the classes for a pie chart and apart from that pie charts are very powerful in demonstrating any dominant category and also single to whole entity relationships. UVW college can use this data to understand which workclasses have more job prospects and can be used to market it to the candidates. Since the majority of the portion is for the Private workclass that means we need further sub categorization of it. This subcategorization is done with help of occupation in User Story 7 and plotted in a scatter plot in upcoming sections.



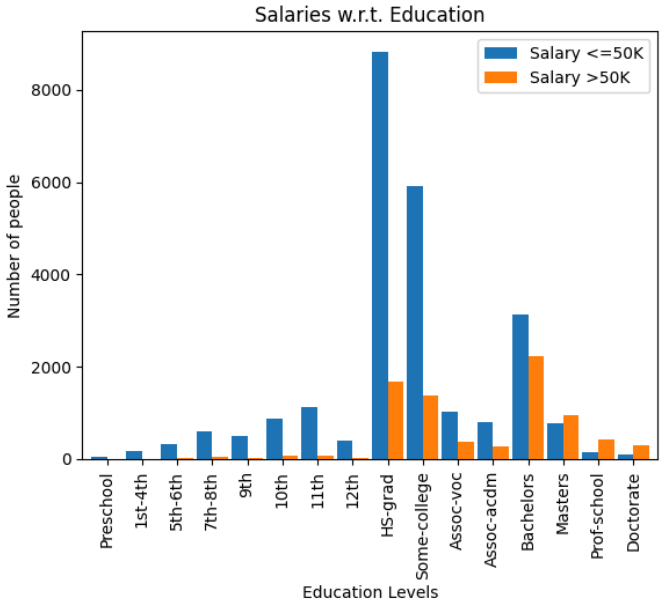
*5. How salaries are varying for the individuals having **native_country** not as USA*

Most universities admit candidates from across the world. In the selected preprocessed dataset most of the data points have the native country as USA but if UVW college is interested in knowing where they can market the degree programs outside the US then they can refer to the insights in the below diagram. We went for geopandas choropleth diagrams as they convey the geographical insights in an easy to understand manner. Clearly we can see outside the USA, most individuals are from Canada, Mexico, India, China and parts of Europe. In order to arrive at this insight we have to do mapping to close to 10 countries as their names were either having "-" or not at all available in the geopandas library e.g. hong, Scotland, etc. We did the mapping of all such countries and grouped the data using pandas while excluding the USA as the native country.



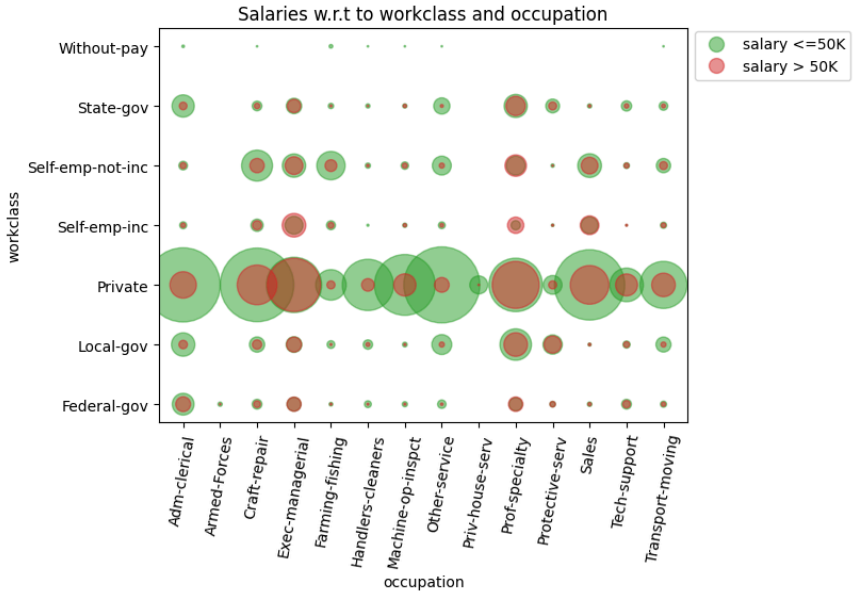
*6. How **education** and **education_num** linked with salaries.*

For coming up with this insight we used education_num, education along with salaries. We group the data by salaries, education and education_num. Education_num is later used to sort the data from preschool all the way to doctorate and these sorted education_num which refers to education is plotted on the bar graph. Bar graphs are pretty good at displaying the categories alongside with ranking as well. In this insight we wanted education to be ranked, that's where the bar chart came very handy. UVW college can use this insight to market the bachelors' programs where we see the most individuals earning more than 50K.



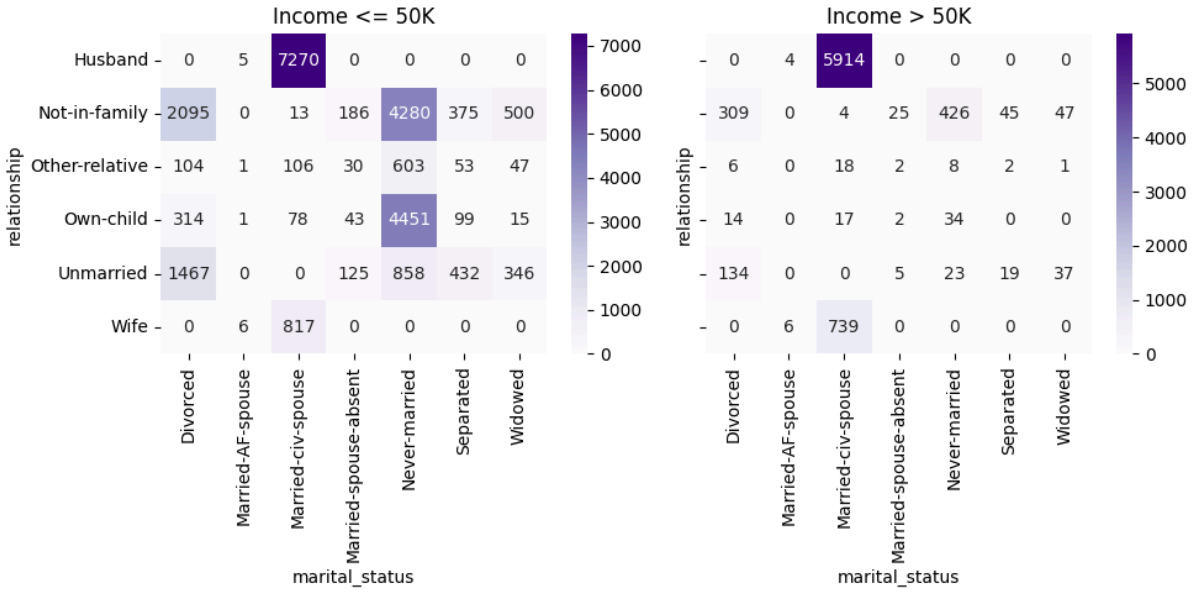
*7. Can **workclass** be used along with **occupation** to determine the salary*

This particular insight is an extension to the user story 4 where majority of the workclass was Private. As part of this insight, we further established the relationship of workclass with the occupation by grouping the complete preprocessed dataset by salaries, workclass and occupation. Size of each of these groups is plotted in the below scatter plot. We went for a scatter plot because they can visualize trends in an easy to understand manner along with the exploration of outliers. Biggest red circle is for the Private workclass in the exec-managerial occupation. UVW can utilize this insight to market their leadership and management degree programs to new students.



*8. How **relationship** and **marital_status** can tell a lot about the individuals salary*

For this last user story, we group the dataset by the relationship and marital_status along with the income to generate insights which subgroup is contributing to the most and the least salaries. We plotted two heat maps, one for less than 50K and another one for more than 50K salary. Further in each plot we can see the relationship between two variables: relationship and marital_status. There are a lot more data points on the less than 50K salary but for the second and third highest number on the left heat graph UVW college can target individuals who are never married for the degree programs that can help them to secure a good paying job and personal relationships down the line.



IV. Challenges Encountered/Questions Arosed

In our analysis we encountered multiple issues which are reported earlier in the progress report and mentioned below as well.

1) *Missing data*: In our dataset we were having missing values represented by "?" so we replaced them with python None.
2) *Missing column headers*: Dataset was missing column headers and we were accessing columns via indexes which made it bit hard to write code so added the column headers
3) *Missing countries in geopandas*: Some of native_countries in the dataset were not available in the geopandas list of countries e.g. Hong, Scotland, etc. so mapped them to correct countries.
4) *Datatype mismatch*: Pandas was reading the dataset as string and certain operations like sorting on education_num was not working as expected so performed the type casting.
5) *Multiple alternatives visualizations*: Coming up with the optimal visualization for the user stories was a bit tricky as we had to plot and compare the visualization before settling on any.

V. Conclusion/ Future Scope

As part of this analysis task, we build a lot of powerful insights that can help the UVW college to market their degree programs while keeping income as the key attribute. From the chosen dataset we were able to utilize the **9 out of 14** attributes except the following attributes - *fnlwgt, race, captial_gain, capital_loss, hours_per_week*. We built a couple of parallel coordinates as well but they were a bit overcrowded and not able to communicate any story well so excluded them from current scope but they are very powerful. Similarly we built a couple of race pie charts but removed them completely to keep the report neutral and unbiased to any race. In the limited time we got to build this analysis we only scratched the surface, and there is a whole world outside to build a lot many powerful insights using data visualization techniques we learned in this course.

Acknowledgement/References

We used a lot of interesting libraries as part of this analysis task. Some of these libraries are for data acquisition, data wrangling while some others are for building insights from the data. All the work is publicly available on Github [7] along with high-definition visualization and complete working code.

[1] https://docs.jupyter.org/en/latest/
[2] https://pandas.pydata.org/docs/
[3] https://matplotlib.org/stable/users/index.html
[4] https://seaborn.pydata.org/tutorial.html
[5] https://geopandas.org/en/stable/docs.html
[6] https://www.statsmodels.org/stable/user-guide.html
[7] https://github.com/asthanaharshita/DegreeEnrollmentMarketing