The Pay/Wage Gap

Final Milestone: White Paper

Madeleine Sharp

DSC680

Capstone Project 1

Business Problem:

Addressing a company's pay gap related to gender (and race) is about more than just money – it is just as much about equity. Given the current socio-political climate, completing equity analyses to close that gap are at the forefront of myriad organizations' minds; after all, closing this gap can bolster business performance, it benefits those who need it most, and it can strengthen the global economy (Lyons, 2019). As our world moves closer to equitable solutions, closing the wage gap is imperative for keeping and maintaining strong employees who offer culture add and much-needed perspectives within the workforce. Happier and more diverse workforces that prioritize equitable treatment and inclusion experience increased productivity, adaptability, and balance (Lyons, 2019).

Background/History:

Historically, in the United States, the term "gender wage gap" (also known as the "gender pay gap") has referred to the disparity in the incomes earned by men and women (Daugherty, 2022; Gould, Schieder, & Geier, 2016). More specifically, this gap refers to that disparity relative to men and women doing the same work/same type of work/same role. Not only has the wage gap impacted people based upon their gender, but so too has it wreaked havoc on people because of their race.

It was not until 1963, in which the Equal Pay Act was passed, that the United States Congress took any sort of major or relevant action to address the gender wage gap in the U.S. (Daugherty, 2022). While the passage of that act is the hallmark of Congress's actions to endeavor for and contribute to a remedy for this social justice issue, it should be noted that the "Equal Pay for Equal Work" movement began in the 1860s (Daugherty, 2022). This is indicative of the amount of time it took between a social movement demanding equal pay for equal work, and subsequent action on the part of Congress – essentially 100 years later.

Given this fact, implementing real change for people is long overdue, and in order to truly identify all areas that impact one's wage (outside of their specific role and/or work tasks), it is imperative to delve into the data to ascertain what additional patterns and discoveries may be found. Along these lines, it is critical to note that not only might a person's gender or race have bearing upon their earnings and income, but other factors may be at play. Organizations need to be aware of the intersectionality of this topic area and should regularly investigate this subject for continual improvement and progression.

Project Overview:

For the scope of my project, I will explore the data surrounding pay/wage discrepancies – particularly taking into consideration any potentially confounding variables outside of gender and race (yet still including gender and race in my analysis) such as seniority, geographic location, years at a company, etc. I will assess this data from an exploratory data analysis standpoint, as well as assessing correlations and relationships, and, finally, linear and multivariate regressions.

Please note: This project is a modified version of my original project plan; please refer to the Appendix (Appendix A) for more information.

Research Questions:

The areas I will be exploring relative to this topic are encompassed within my research questions, which include:

1. What does the current income landscape look like with respect to:
    a. Gender?
    b. Marital Status/Role in Household?
    c. Race?
    d. Ethnicity?
    e. Age?
    f. Educational Level?
2. What does the current income landscape for data professionals look like with respect to:
    a. Gender?
    b. Job Title?
    c. Country?
    d. Industry/Sector?
    e. Years of Experience?
    f. Management Level?
3. How do these two areas compare? What similarities or differences exist?

Data Explanation – Data Preparation & Data Dictionary:

For this project, I used two main datasets to answer my research questions. These datasets and their descriptions are below.

2020 U.S. Census Data

This data comes from the U.S. Census Bureau's Table 1A data (Educational Attainment, People 18 Years Old and Over by Total Money Earnings in 2020 (including Total (All) Work Experience, Age, Race, Hispanic Origin, and Sex (aggregate)). (https://www.census.gov/data/tables/2020/demo/educational-attainment/cps-detailed-tables.html)

This data is in the form of summary (or aggregate) data, so I will not be assessing individual incomes for individual people with this data, but rather summary income data for cohorts/groups of people based upon my research question interests.

2020 Data Professional Salary Survey Data

This is an interesting dataset that includes income data for data professionals. This dataset includes individual (not aggregate) data and comes from The 2020 Data Professional Salary Survey, administered by Brent Ozar Unlimited (https://www.brentozar.com/archive/2020/01/the-2020-data-professional-salary-survey-results-are-in/).

This data will be used to make an assessment relative to my second research question (and subsequently, my third question).

Overall, with respect to these data selections, I wanted to utilize a combination of individual, aggregate, and up-to-date data that was relevant. Some datasets I explored did not have actual salary values and only had two clusters of salary value groupings; I wanted actual salary values (even if it meant a measure of central tendency) versus merely a classification of all salaries into one of two groups. Given those desires, I ascertained that the best datasets for use for my project would be those outlined above.

To prepare my data for analysis usage, I complete the following data preparation steps:

- Pull in the data
- Read/view the data
- Assess for missing values
- Assess for unique or odd variables
- Assess for data types; change any necessary
- Select which variables are of interest to me and the project topic
- Transform and slice the data based upon those variables

Please note: This project and its accompanying data is a modified version of my original project and data sourcing plan; please refer to the Appendix (Appendix A) for more information.

Additionally, a description of all data variables is included in the Appendix of this document (Appendix B); please refer there for more information.

Methods:

The analysis methods utilized for this project include the following below. The primary focus of my analysis is to assess the relationships between income and other demographic variables in an endeavor to determine what might be related to income earned, and what might therefore explain any wage gaps.

Prior to any analyses, the data was sourced, pulled in, and transformed/cleaned as needed.

- EDA (exploratory data analysis):
    - Distributions and Bar Charts: To assess wages/income earned relative to the other variables of interest.
    - Descriptive Statistics: To grant an overall view of the data from central tendency and variance standpoints.
    - Correlation Heatmaps: To assess the relationships between variables and their respective correlation coefficients.
    - A combination of these methods will be utilized for exploring all of my data.
    - In order to be utilized in EDA, certain categorical variables have been converted to dummy variables; this also accounts for any multi-collinearity that may exist.
- Linear & Multivariate Regression (for 2020 Data Professional Survey Data only):

- I will use linear regression to assess various relationships that I would like to investigate more deeply post-EDA.
- Again, certain categorical variables have been converted to dummy variables; this accounts for any multi-collinearity that may exist.
- The relationships that I will assess here are:
    - Gender (IV)* and Salary (DV)* - linear
    - Management Level (IV) and Salary (DV) - linear
    - The interaction of Gender (IV), Management Level (IV), and Salary (DV) – multivariate

My reasoning for my methodology choices includes the below:

- Distributions and bar charts are a standard way to view and assess data related to populations; given this, these types of EDA were fitting for my data.

- Descriptive statistics are always useful for granting some "overview insight" into the data one is working with.

- Correlations/correlation heatmaps are useful for assessing relationships between multiple variables and how they impact one another – this was useful for my assessment of how variables were related to income or how they might explain pay gaps.

- Regressions are useful for assessing how an independent variable (or more than one IV) impact a dependent variable – in this case, the variable of income.

*__*Please note*__: IV stands for Independent Variable and DV stands for Dependent Variable. The DV is the variable that we are assessing to what degree it is or is not dependent upon the IVs.

Analysis:

The results of my analysis are below, and describe the results from my EDA, as well as from my regression analyses. All visualizations as a result of analysis, including tables, figures, etc., are included in the Appendix (please see Appendix C.1).

EDA:

Distributions (Histograms) and Bar Charts:

From the EDA with distributions and bar charts, the following findings have been gleaned:

- **2020 U.S. Census Data:**
    - Gender and Households:
        - Married couples were a smaller subset of this data, but overall had the greatest median income.
            - This makes sense, given that married households often have two people bringing in income.

        - There are more female householders with no spouse present in the data, but they make quite a bit less than men householders with no spouse, from a median income standpoint.
        - Additionally, female householders make up more of the dataset than male householders, yet still make less median income-wise than their male householder counterparts.
        - Overall, female-headed and non-family householders made the least median income, whereas households with more than one person present made the most median income.
    - Race and Ethnicity:
        - Non-Hispanic white individuals comprise the majority of the dataset, but do not make the most median income overall.
        - Asian individuals make up the smallest portion of this sample population, however they make the greatest median income overall.
        - Black individuals also make up a small portion of the dataset, and they also make the least median income overall.
        - Slightly more of the dataset is made up of Hispanic individuals of any race, and they make only slightly more median income overall than black individuals do.
    - Age:
        - The majority of the age cohort is comprised of those 65 years and older. However, these people make the least overall median income from the dataset (in line with those in the 15-24 age cohort).
        - 15-24 year-olds comprise the least amount of people in this dataset, and also garner the least overall median income (in line with those 65 and older).
        - Those in the 45-54 age category make the most overall median income.
            - This is likely when someone is at the "peak" of their career. (It is important to note that this could be different for different people within this category).
    - Education:
        - The population of education level groupings as well as the median income earnings tend to be the most in line, out of all of the graphs.
        - The fewest represented are those with only a HS degree and no college - they also make the least median income overall.
        - The most represented are those with a BS degree or higher - they also make the most median income overall.

Overall, from these bar chart analyses and distributions, the following should be considered:

- It is important to note that these groups are not mutually exclusive. For example, there could be a person who is part of a married household but is also considered the female

householder. As another example, a female householder present within the data could be a non-white Hispanic. Etc.

- In this same vein, it is important to note that even though a certain cohort may show a median picture for that cohort, there are sub-cohorts that make up that cohort that may experience a difference. For example, you could have a black woman with a BS degree who earns less than some of her counterparts.
- It is important to ask questions about survey response relative to culture, access, literacy, etc.

  * Are certain groups of people more likely to fill out the Census survey?

  * Are there groups of people who do not have adequate resources to complete the survey, or who are not familiar with how to do so?

  * Culturally, are certain people more inclined or not inclined to complete the survey?

- **2020 Data Professionals Salary Survey Data:**
  - Gender:
    - This dataset is primarily comprised of male responders relative to the other gender categories.
    - A good portion (combined) of respondents did not prefer to respond, responded with something silly (NA), or were not asked on their survey.
    - A very small portion of this survey included data from female respondents as well as non-binary/third gender respondents.
    - The second bar chart shows the greatest earnings concentrated in the non-binary/third gender category, but this is likely due to skew, given how few respondents existed within that category of the population.
    - Additionally, from the second bar chart, it appears that females made more than their male counterparts, despite making up a small portion of the survey sample, but again, this could be related to skew.
    - Quite a bit of gender information is missing from this, so it is tough to know what true gender category those respondents would fall into.
    - A few outliers exist, so that will need to be taken into consideration with any assessment/conclusion.
  - Country:
    - The greatest salary earnings came from those living in Hong Kong.
    - The next greatest salary earnings came from those living in Switzerland (2), Bermuda (3), the U.S. (4), and Uganda (5).
    - The U.S. still made the top 5 in terms of salary earnings.
    - Not certain how much some of these categories and the income is related to skew.
  - Job Title, Employment Sector, Gender:
    - Male and Not Asked were the overwhelming majority of the data representation for job titles.
    - More males than females are represented in most of the roles, but this could be due to the dataset imbalance of the gender classes.

        - Most work in the Federal Government Sector.
        - Sales garner the most income of all of the positions (it is common for sales to have high income).
    - Management, Years of Experience
        - More men are managers than women, but again, this could be due to the gender class imbalance of this dataset.
        - There are fewer managers than not, but managers make more in salary.
        - Most of the dataset representation has 20 years or less of experience within the specific role.
        - Over time, we can see a slight increase in salary (for the most part) as years of experience within the role go up.
        - The representation by years of experience by gender takes on a right-tailed distribution.

Descriptive Statistics:

Descriptive statistics were assessed throughout various sections of my code file, so within my code, they did not have their own reserved section. However, I describe them here.

2020 U.S. Census Data:

For this dataset, the descriptive statistics run on the data were indicative of and reflective of the groups above, but as a whole overall. For example, instead of viewing the population and median income per grouping within a category (age, race, gender, etc.), the descriptive statistics granted us information about the entire category itself for both population and median income.

Of course, for my analysis, viewing the descriptive statistics for the categorical wholes was not as informative and telling with respect to my research question as were the above analyses.

2020 Data Professionals Salary Survey Data:

For this dataset, the descriptive statistics run on the data only return output for numeric values (which is standard). These descriptive statistics look at the data from an aggregate standpoint. From all of the people in the dataset, the average (mean) salary was 93,946.16. For the "years with this type of job experience" the mean value was 7.5 years.

Overall, while descriptive statistics can be helpful for describing data in general, in my specific project instance, they did not necessarily contribute to my research question answers, but I felt they were important to include as supplementary analyses.

Please note: To review the descriptive statistics output tables, please refer to Appendix C.2.

Correlation Heatmaps:

From the EDA with correlation heatmaps, the following findings have been gleaned (correlation heatmaps were only used for the 2020 Census Data):

- **2020 U.S. Census Data:**
    - Gender and Households:

        - A moderately strong, positive correlation exists between the number of people in a category and the median income earned (0.46) (makes sense - with more people is more money added to the "pool" of money from which the median in determined).
        - A somewhat strong, positive correlation exists between being a married couple and your median income (0.69), which we could see from the EDA bar chart of this variable.
        - A moderately strong, positive correlation exists between the variables of median income and family households (0.43).
        - A moderately strong, negative correlation exists between the variables of median income and being a female householder - this makes sense given our bar chart EDA; as one variable increases (being a female householder) the likelihood of the other decreases (median income) (-0.46).
    - Race and Ethnicity:
        - A strong, positive correlation exists between being Asian and the median income earned (0.83) This makes sense, given what we saw in our EDA bar chart.
        - A strong, negative correlation exists between being black and your median income (-0.67), which we could also see from the EDA bar chart of this variable.
    - Age:
        - Strong, negative correlations exist between median income and the 15-24 age cohort (-0.57), as well as between median income and the 65 years and older cohort (-0.59). This makes sense, given our EDA of these variables. It appears that overall, young people and elderly people tend to make less.
        - Moderately strong, positive correlations exist between the variables of median income and the 35-44 cohort (0.43) as well as the 45-54 cohort (0.55).
    - Education:
        - A strong, negative correlation exists between median income and those with only a HS degree (-0.71). This was also deduced based upon the EDA bar chart.
        - A strong, positive correlation exists between the variables of median income and having a BS degree or higher (0.96). We could ascertain this from the bar chart EDA.

Overall, the findings from the correlation heatmaps with these variables are indicative and reflective of the EDA histogram and bar chart findings - they affirm what I explored earlier.

Please note: To review the descriptive statistics output tables, please refer to Appendix C.3.

Regression:

For this final step, and the main purpose of what I wished to conduct my project surrounding this topic, was to assess just how influential certain characteristic variables might be on someone's earnings.

- Linear Regression
    - Gender and Salary
        - Most of the gender groupings are negatively correlated with salary.
        - Non-binary/third gender is the exception, but very few of these people existed within the dataset.
    - Management Level and Salary
        - The only truly significant p-value is for the gender grouping "Not Asked" - at 0.01.
- Multivariate Regression
    - Gender, Management Level, and Salary
        - The combination of these variables did not drastically change which or how they had an effect on salary.

Conclusion:

From my overall analysis, I conclude that more robust analyses may need to take place.

In the preliminary EDA, it is apparent that certain pay imbalances/income gaps still exist – particularly when it comes to gender, race, household type, education, and age. From a data professionals standpoint, overall, certain jobs or sectors prove to be more fruitful. The regression analysis here also appeared somewhat inconclusive.

However, even though these observations are present, I came to find that the datasets I used were not the most robust for the purposes of using them to assess or predict relationships.

I think it would prove useful and behoove those who are researching in this space to capitalize upon this via the incorporation of more data from more sources – also taking into account different types of data. For my project, one dataset is U.S. focused, while the other is global. Both have similar variables, but one is a population subset with the other being an overarching population aggregate view.

Assumptions:

From the beginning of this project, I had assumptions about what I would find within this data. Some of these included:

- A pay gap still exists between men and women and between various racial groups. What I did not assume or expect to find was that from the U.S. Census dataset, that Asian individuals would out-earn all of the other groups (I assumed white individuals would).
- More men would be represented within the data professionals survey data – I found this to be true, although, I am not sure how accurate the representation is (did less women just fill out the survey, or are there truly less women in the data field?).

Limitations:

With respect to the limitations of this project, the following are considerations to be taken into account:

Personal Limitations

- As a single individual, I am limited in the amount of work I can complete on this topic, and the scope of which that reaches.
- I am limited by the data available for me to analyze, and even then, a smaller subset of that, since I was not able to analyze all data available on this topic for my project and within the project timeline.
- While my undergraduate experience focused on psychology, sociology, and women & gender studies, I still am most certainly not an expert to the fullest extent in all of these areas.

Dataset Limitations

- Overall, my data included a good mix, but I think each source was limited in its scope. This project could be expanded upon with additional data sources and types.
- Certain data that I wanted at an individual level I could only find in aggregate form.
- Some of my data included imbalanced classes – in the future, if I choose to run a ML model on the data, oversampling methods like SMOTE would likely need to be employe.

Challenges:

Some of the challenges I faced during this project include:

- Finding a strong, reliable data source that had enough information for me to complete my analyses on, without being overly complicated.
    - Additionally, selecting the variables from the dataset was challenging in its own right.
- Selecting the best EDA and analyses/models to run was tough, especially when using both categorical and numerical features/variables.

Future Uses/Additional Applications:

Future uses of these datasets as well as the findings may involve additional research in this topic area. Additionally, from an applicability standpoint, the findings I have come across may grant additional support to and for the case regarding the pay wage gap in the United States. This may be important for researchers, policy makers, social scientists, employers, DE&I efforts, etc.

Recommendations:

My recommendations moving forward are as follows:

- Look into additional data sources and types from various institutions.
- Run more robust regressions, as well as other predictive model types.
- Complete a more thorough assessment of the current literature, as well as differences by industry type, job title, city, etc.
- Slice/dice/look at the data in different ways and from different angles.

Implementation Plan:

Following the culmination of this project, my proposed implementation plan would be:

- Each company/organization should look into its own potential income gaps or wage biases.
    - Analysis should ensue, and based upon that analysis, certain standards should be implemented and maintained.
    - This should be a regular effort across time.

Ethical Assessment:

For my specific project, the ethical considerations regarding this data/my project include the following:

- Data sources: I need to have enough data from enough sources to be representative of the population and to minimize any potentially biased or skewed data.
- Data collection: I also need to be aware of the ways in which the data was collected – this has the potential to impact outcomes and the meaning of data insights post-analysis.
- Data agenda: Do certain sources of data/collection methods indicate a specific agenda/bias that I should be aware of? How can I ensure the data I use it accurate and was collected with integrity?
- Data impact: Who does this data impact, and what is being done with said data? Are the insights found going to be used to be helpful, or to hinder? Will this data, or has this data, gotten into the hands of those who have the power and/or ability to do something constructive with it? What will my project add to this, or will it?

Appendix:

Appendix A:

Original Project & Data:

For this project, my original plan was to use the following datasets to assess my original project topic. My current project topic is quite similar to my original project topic and is merely a modification of my initial plan. Initially, I wanted to use individual U.S. income data (versus aggregate) and I wanted to be able to make predictions about individual incomes based upon features/variables within that type of dataset.

As I worked through some initial coding and analyses, I ran into some challenges and needed to lean on my contingency plans I had put into place for a more reasonable and successful outcome. Please see below for my notes regarding my original data and project work.

Initial Dataset #1:

This dataset is from the U.S. Census Bureau and is an adult dataset from 1994 (https://archive.ics.uci.edu/ml/datasets/census+income) that has been compiled and housed at UCI (The Machine Learning Group). I found this dataset to be too outdated (and my project plan did not include a time series analysis) and its income data was not in the format that I desired - only two salary groupings exist in this data: equal to or less than 50,000 dollars or greater than 50,000 dollars. These values were not going to be useful to me for what I was interested in exploring.

Initial Dataset #2:

This was set to be my initial primary dataset, but unfortunately, the data and its files proved tough to deal with, or even to pull in. The data itself is sourced from the Panel Study of Income Dynamics (PSID) microdata (https://psidonline.isr.umich.edu/default.aspx), which is collected, housed, and maintained by the University of Michigan's Institute for Social Research and Survey Research.

This specific dataset (PSID) is included as a built-in package in both Python and R (https://pypi.org/project/psid_py/; https://cran.r-project.org/web/packages/psidR/index.html), but these are packages developed by individual developers and some of the methodologies and code for using the data are outdated, given that the PSID website has been updated/changed, and it was not clear that these packages accounted for that.

After attempting multiple formats and methodologies for getting this data to "read in" so that I could finally work on my analyses, I decided going down this route would not be the most realistic or fruitful with the time and experience I possessed. Additionally, this data included 164

variables, and an extensive codebook. I felt that to successfully complete a project using this data, it would require additional skill, time, and expertise.

Appendix B:

Data Dictionary Variables & Dataset Information:

2020 U.S. Census Data

Variables I used from this dataset include:

- Characteristic/Characteristic Values: this was the person characteristic – their race, their gender, their household status, etc. Categorical.
    - I ended up splitting this characteristic column and its respective number (people) and median income data into multiple data frames – one data frame for each topic area (gender/household, race/ethnicity, age, education, etc.)
- Number (people): this was the number of people present within the data collection sample; in other words, those that completed the survey. Numeric.
- Median Income: this was the median income value for each respective characteristic group. Numeric.

2020 Data Professional Salary Survey Data

Variables I used from this dataset include:

- Salary USD: this is the salary a person earns – it is an exact amount, not a range or bracket. Numeric.
- Country: this is the country in which one lives. Categorical.
- Job Title: this is the person's job title; there are a set number of categories a person could elect that best reflected their job title. Categorical.
- Manage Staff: this is a Yes/No question inquiring whether a person is a manager and manages staff. Categorical.
- Years with this Type of Job: this is the number of years a person has had this kind of role. Numeric.
- Population of Largest City within 20 Miles: this is a range/bracket variable of the closest estimate to a person's largest city population within 20 miles. Categorical.
- Employment Sector: this is the industry in which one works; there are a set number of categories a person could elect that best reflected their employment sector. Categorical.
- Gender: this is the person's gender. Categorical.

Please note: All of the data for these variables come from self-report survey data; both for the U.S. Census Survey and the Data Professional Salary Survey.

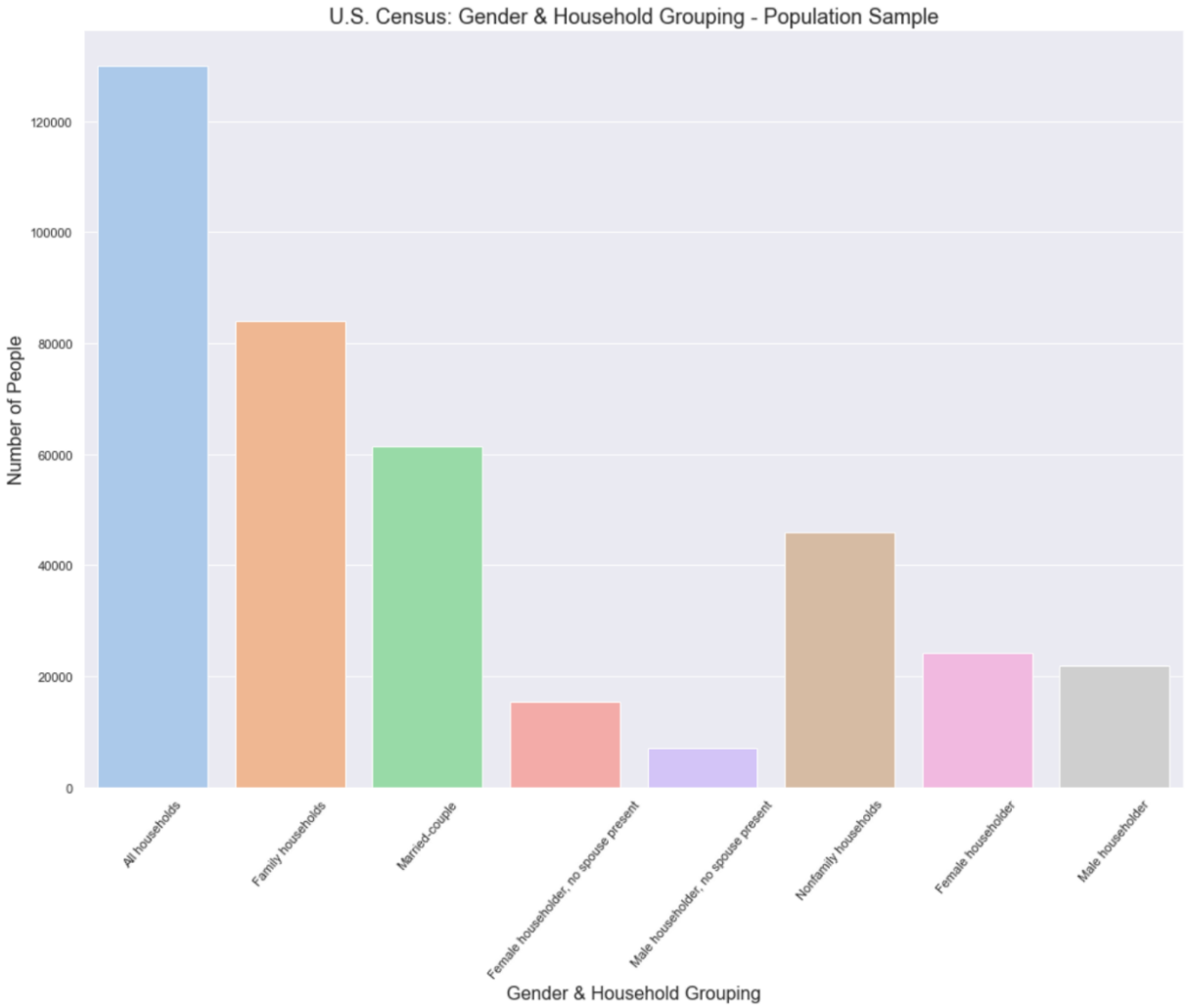
Appendix C:

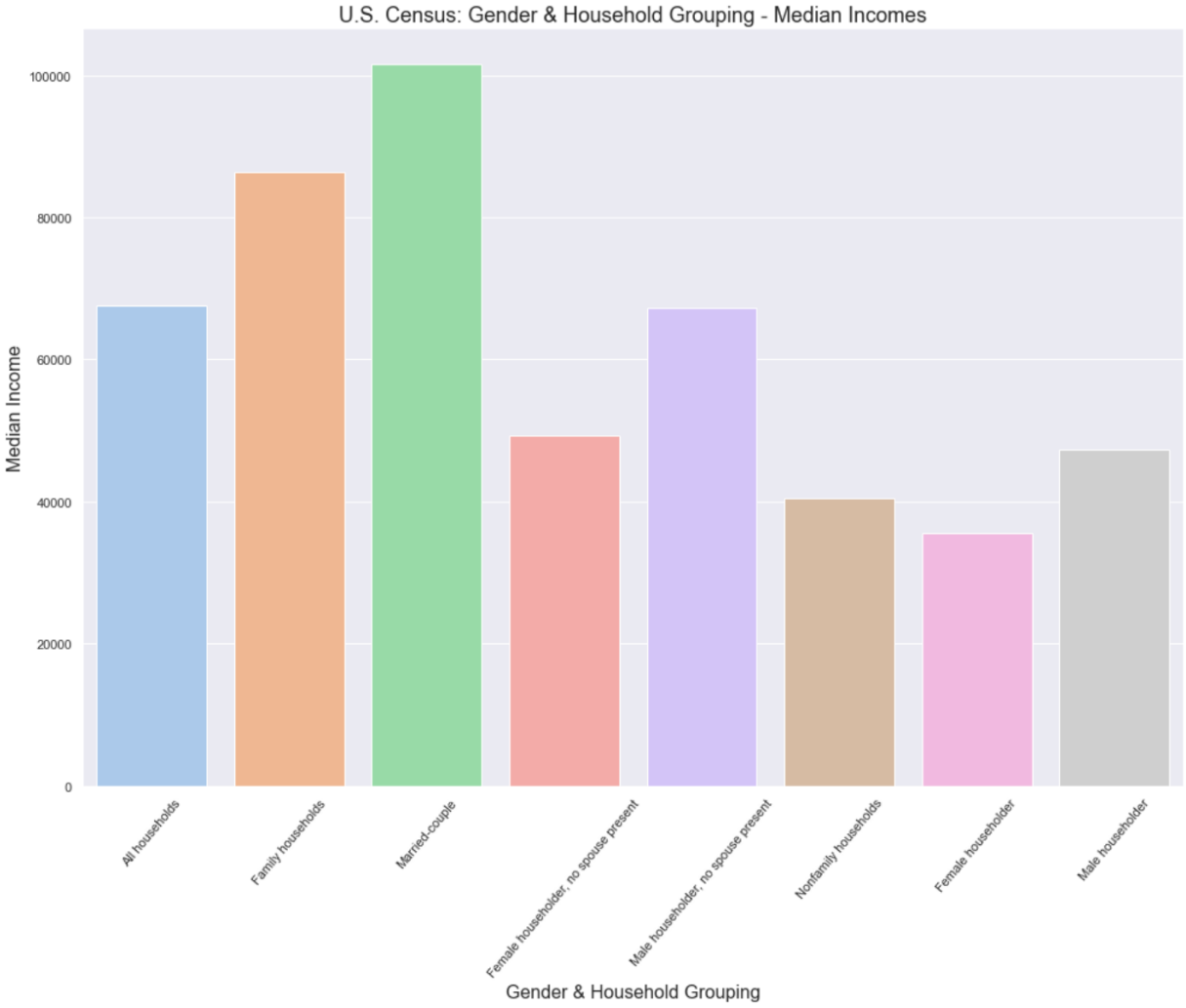
Data Analysis – Accompanying Visualizations:

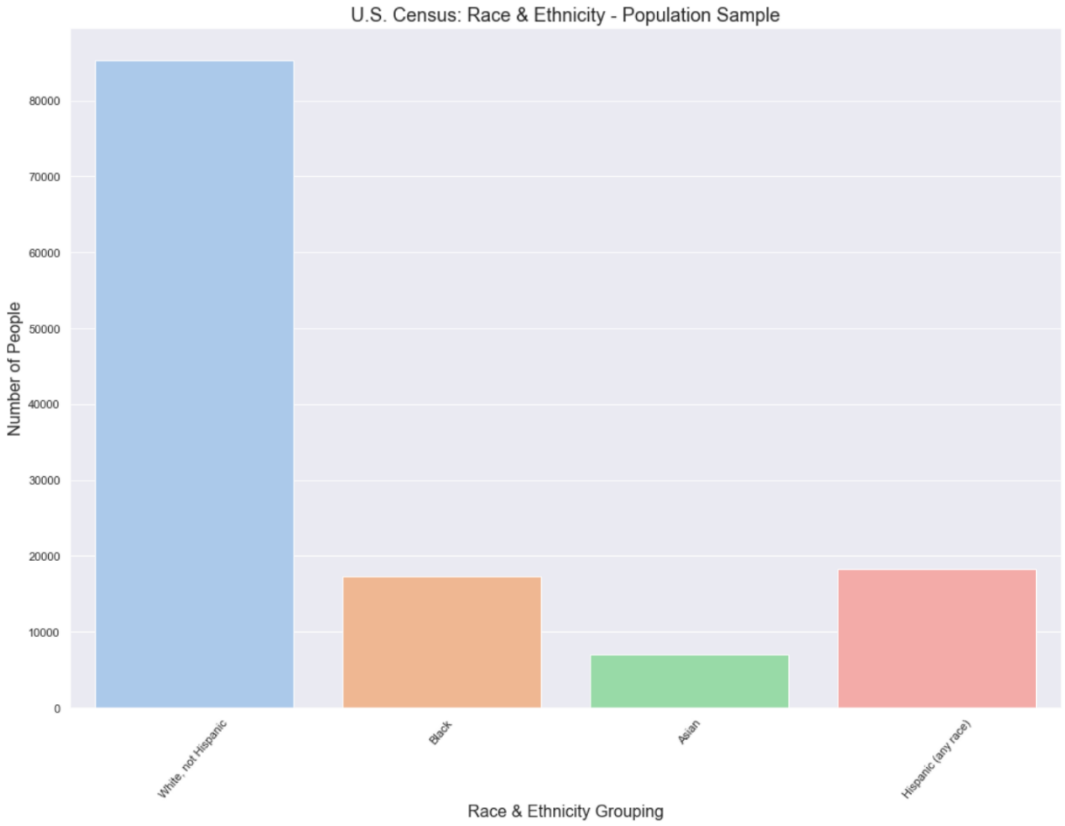
Please see all visualizations below.

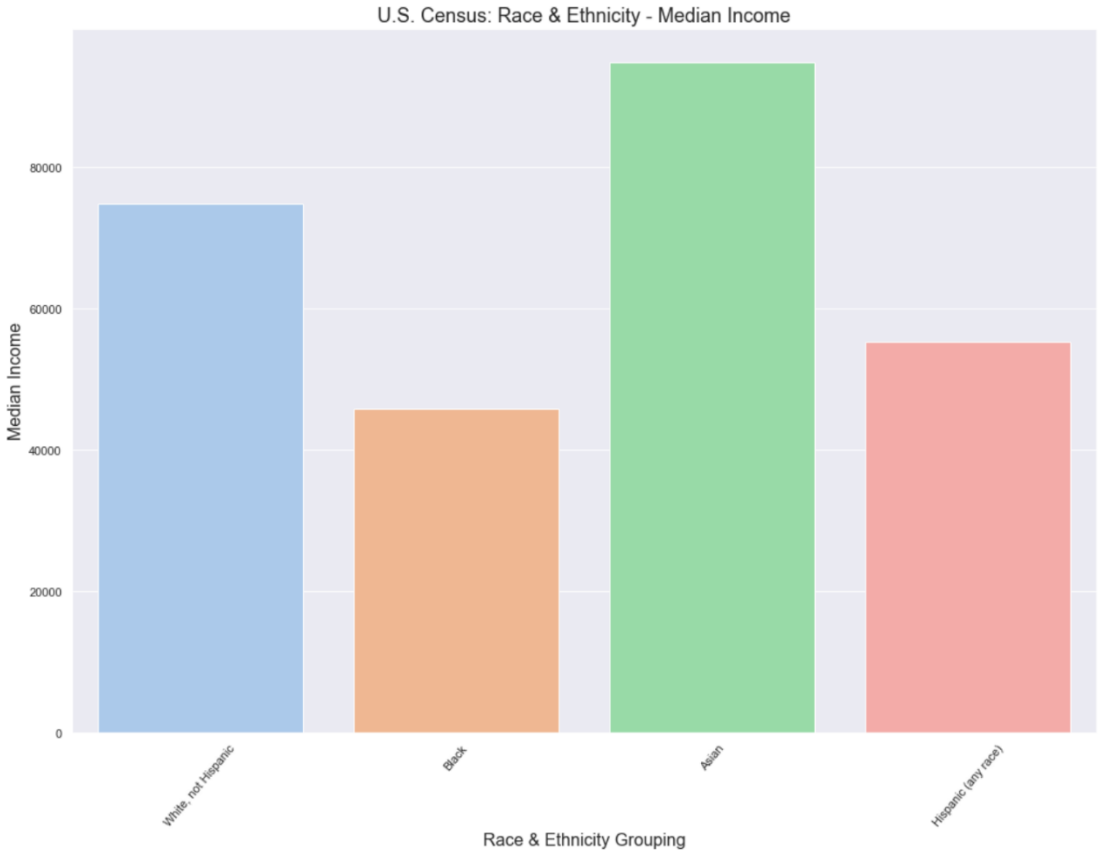
1. ***Distributions (Histograms) and Bar Charts:***

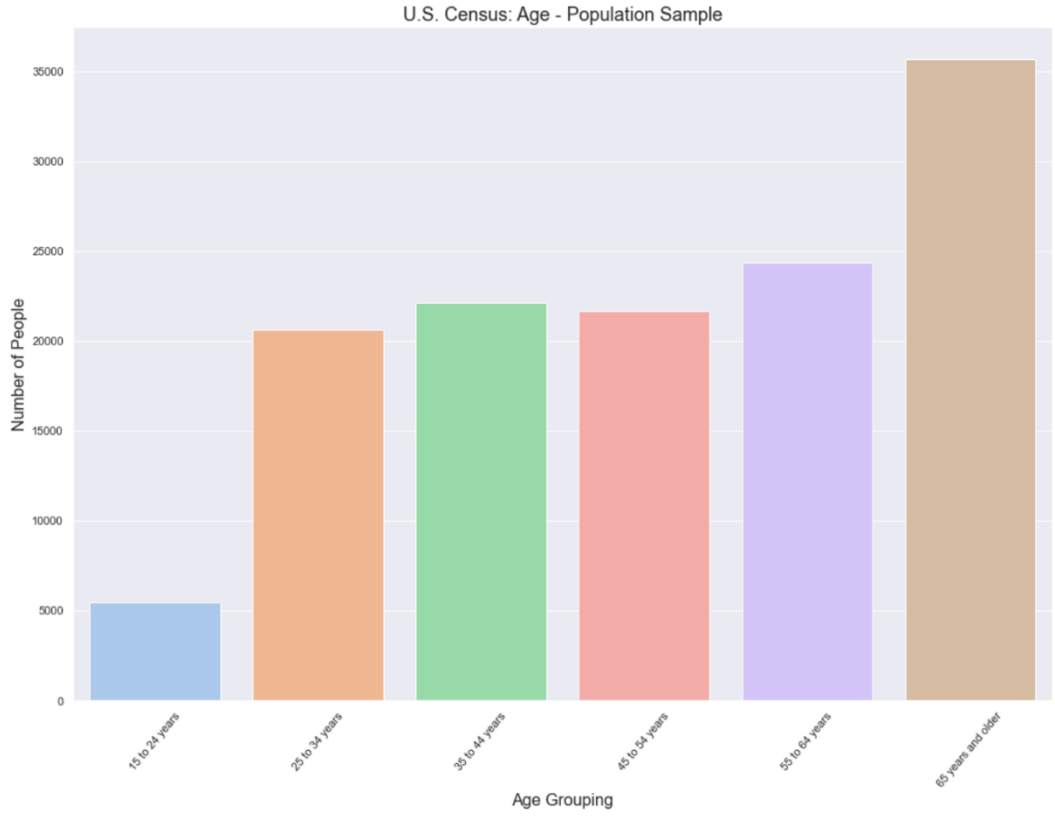
2020 U.S. Census Data:

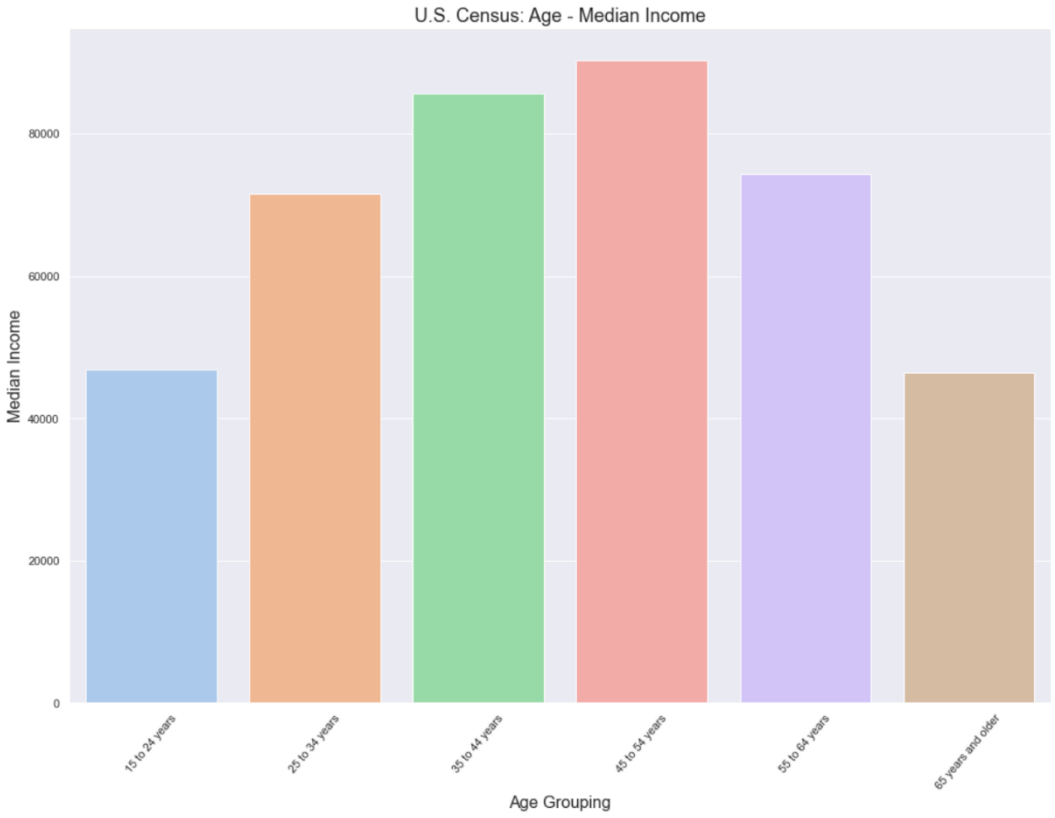


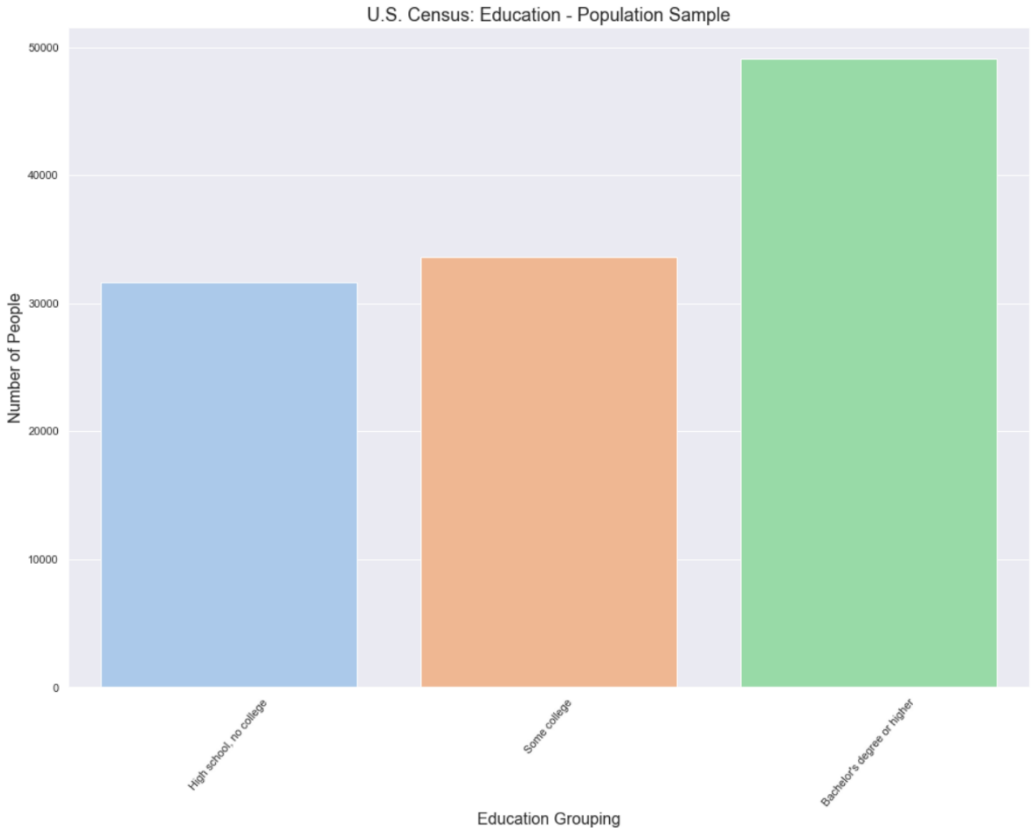

U.S. Census: Gender & Household Grouping - Median Incomes
Median Income
100000
80000
60000
40000
20000
0
All households
Family households
Married-couple
Female householder, no spouse present
Male householder, no spouse present
Nonfamily households
Female householder
Male householder
Gender & Household Grouping

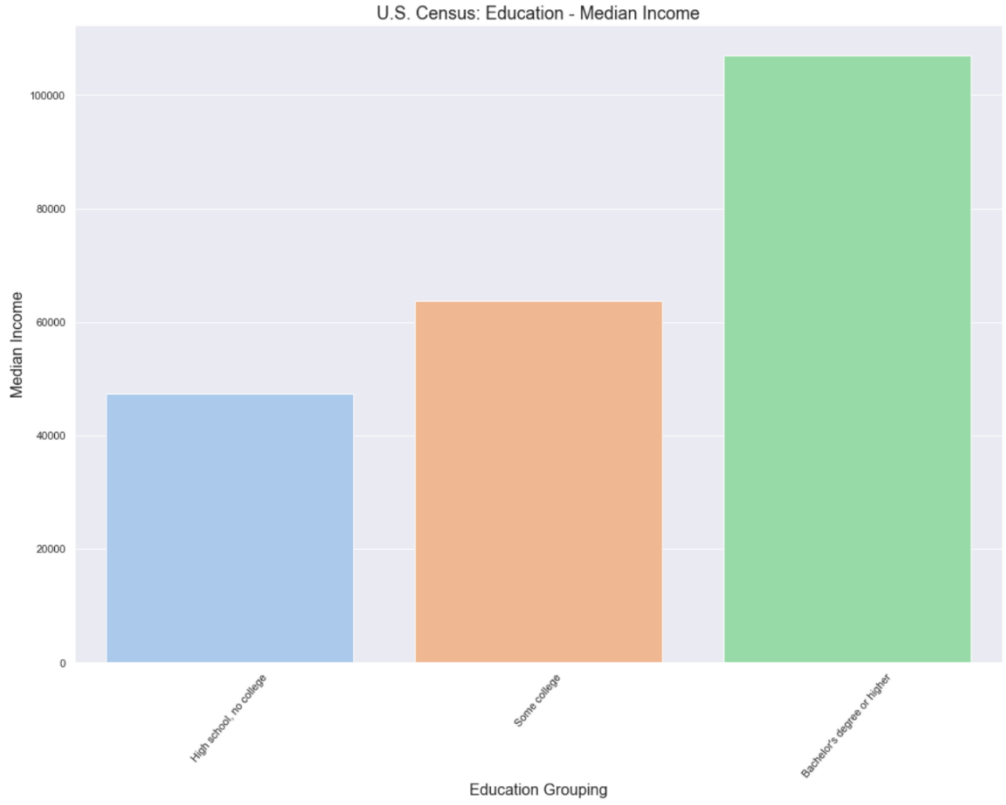

U.S. Census: Race & Ethnicity - Population Sample
Number of People
0
10000
20000
30000
40000
50000
60000
70000
80000
White, not Hispanic
Black
Asian
Hispanic (any race)
Race & Ethnicity Grouping


U.S. Census: Race & Ethnicity - Median Income
Median Income
0
20000
40000
60000
80000
White, not Hispanic
Black
Asian
Hispanic (any race)
Race & Ethnicity Grouping

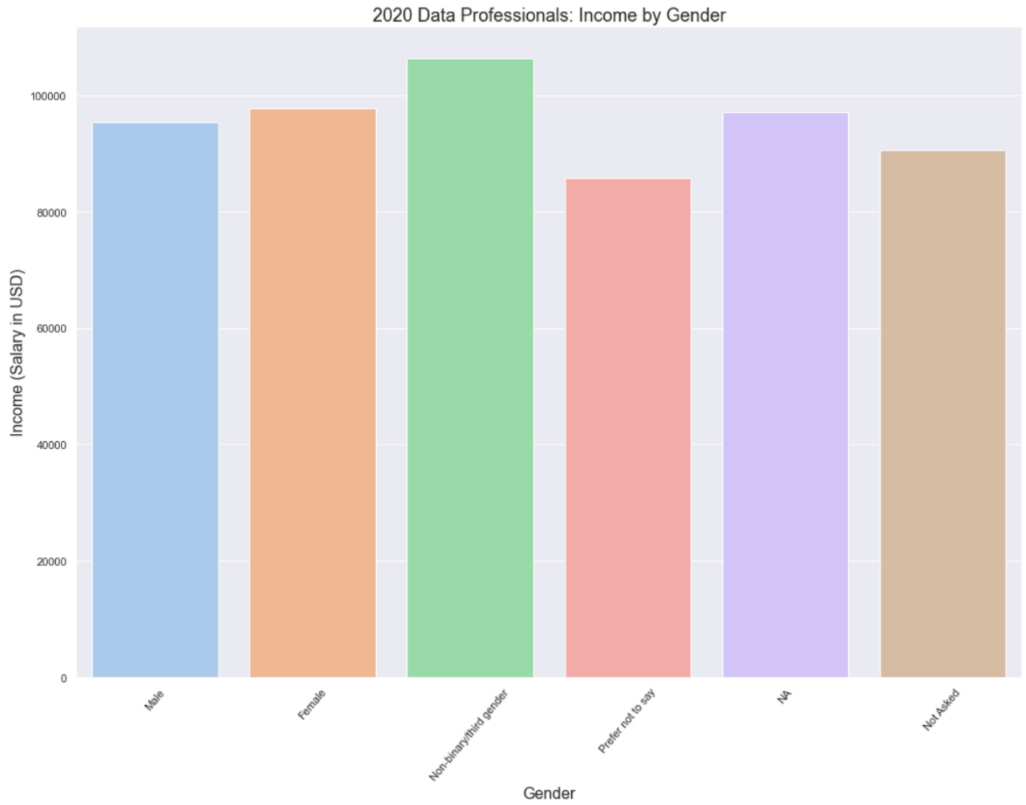

U.S. Census: Age - Population Sample
Number of People
0
5000
10000
15000
20000
25000
30000
35000
15 to 24 years
25 to 34 years
35 to 44 years
45 to 54 years
55 to 64 years
65 years and older
Age Grouping

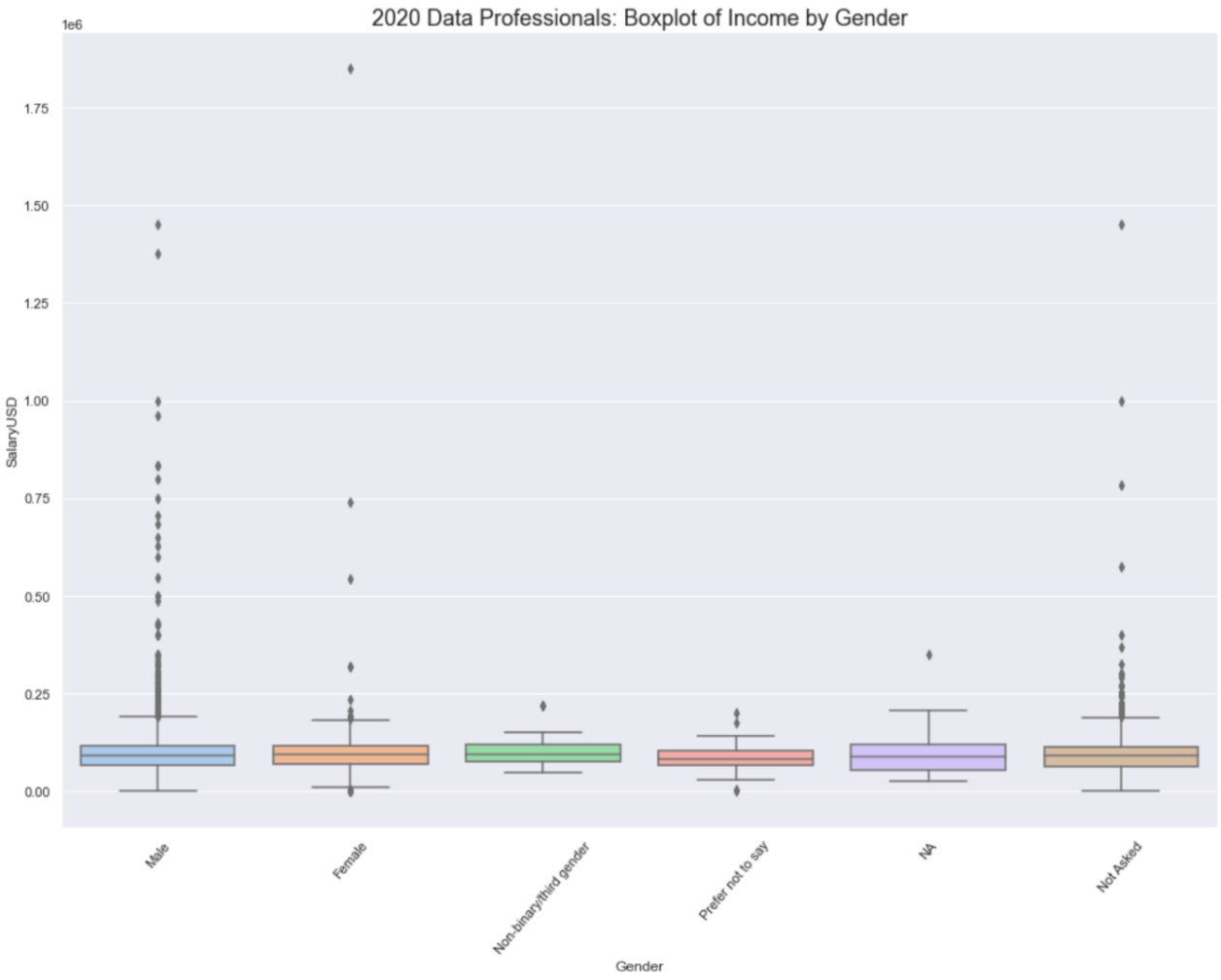

U.S. Census: Age - Median Income
Median Income
0
20000
40000
60000
80000
15 to 24 years
25 to 34 years
35 to 44 years
45 to 54 years
55 to 64 years
65 years and older
Age Grouping

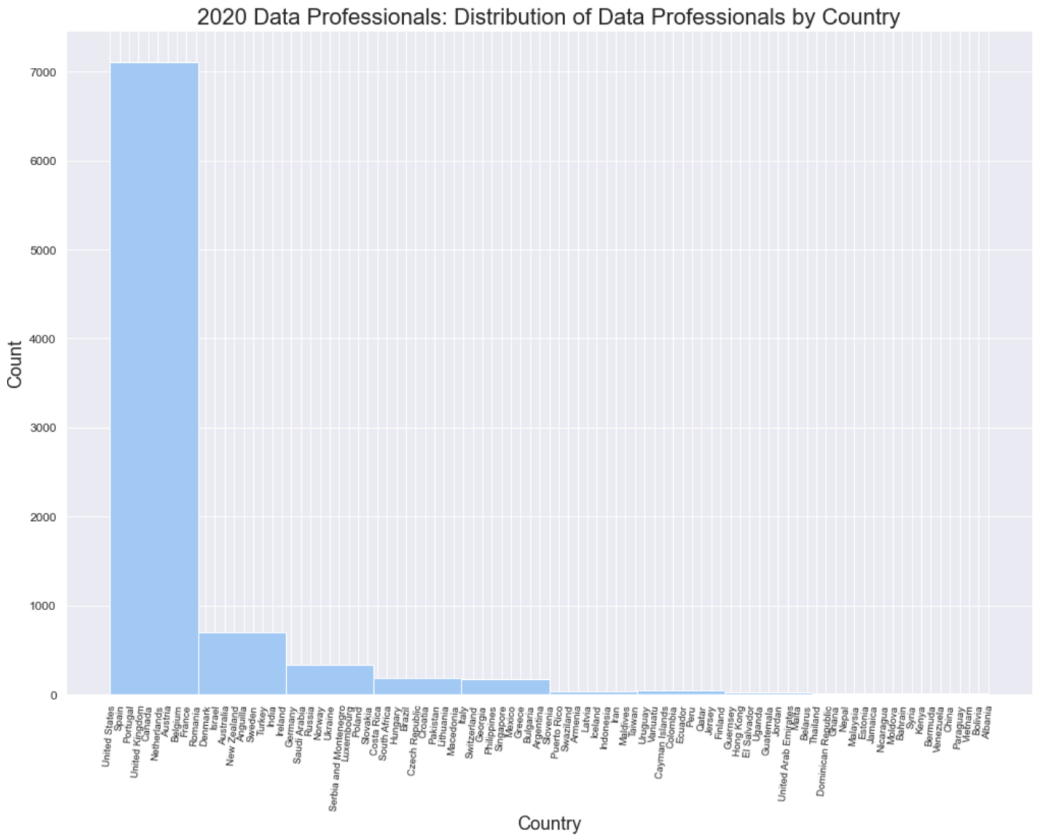

U.S. Census: Education - Population Sample
Number of People
50000
40000
30000
20000
10000
0
High school, no college
Some college
Bachelor's degree or higher
Education Grouping

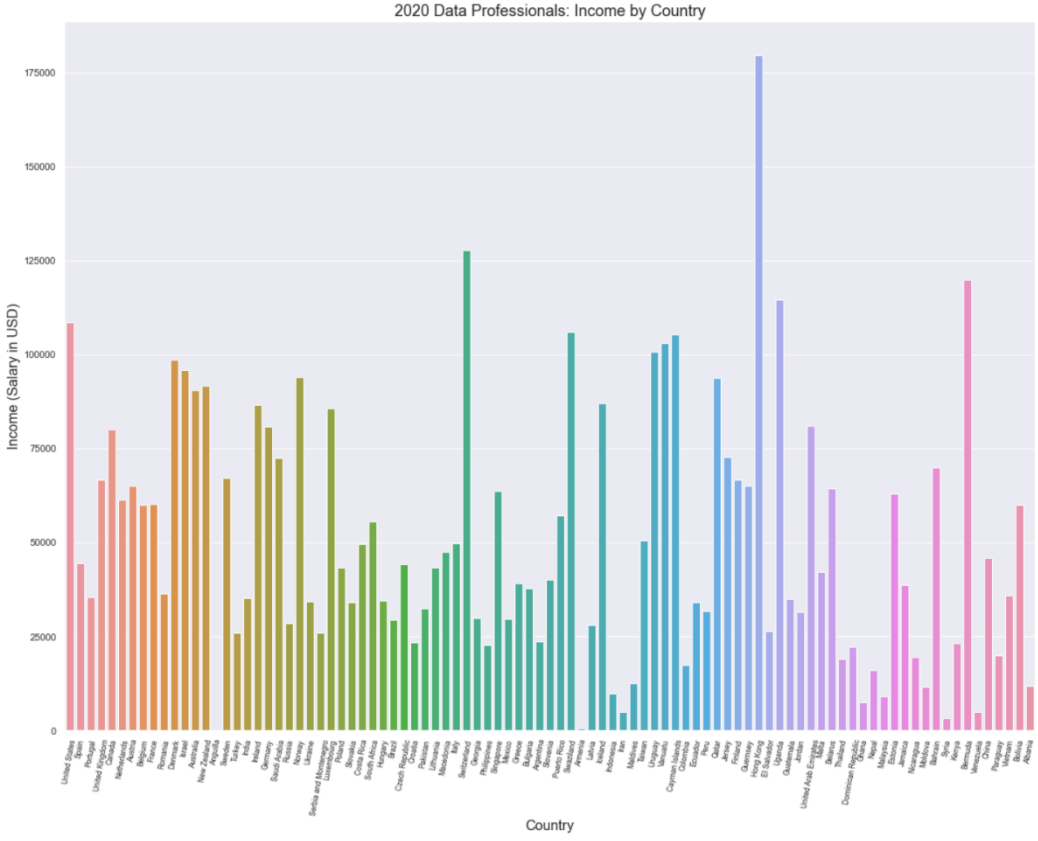

U.S. Census: Education - Median Income
Median Income
100000
80000
60000
40000
20000
0
High school, no college
Some college
Bachelor's degree or higher
Education Grouping

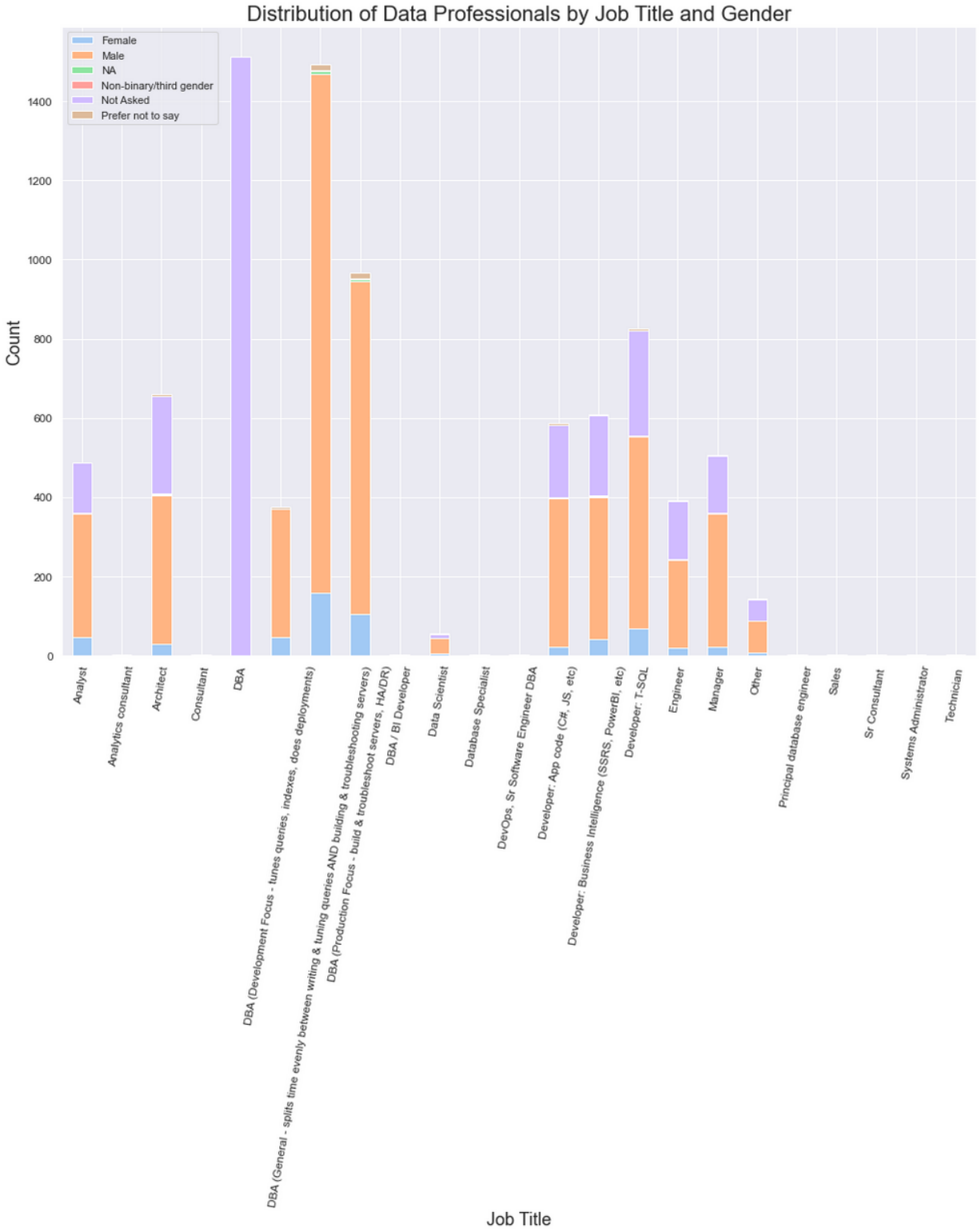
2020 Professionals Salary Survey Data:

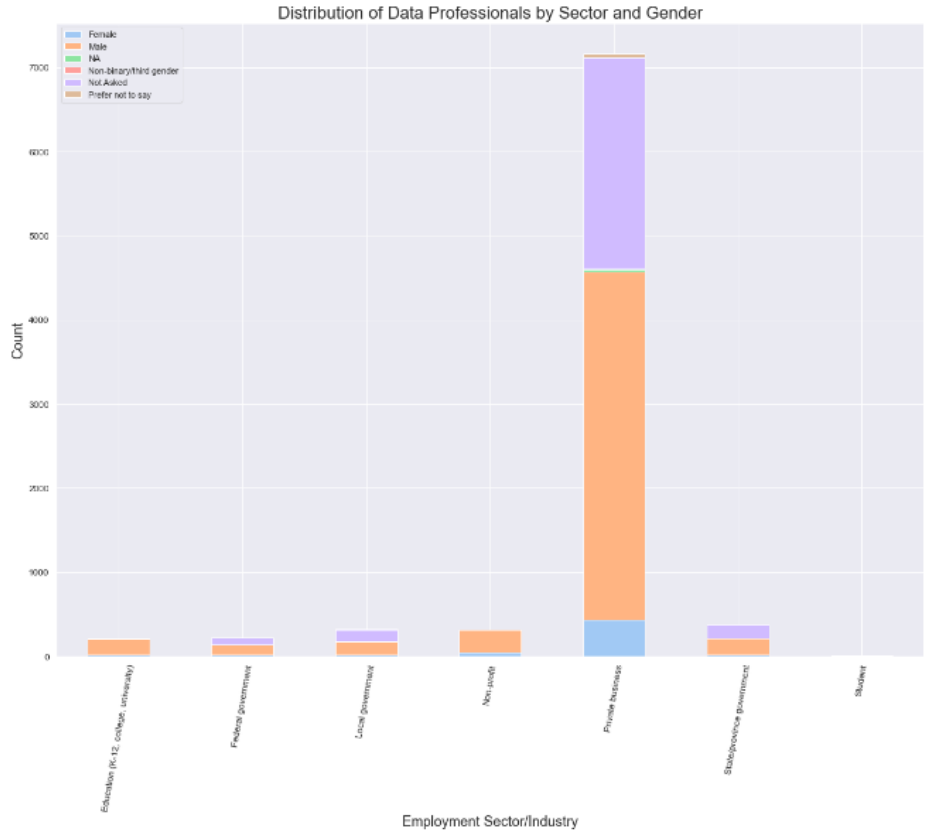


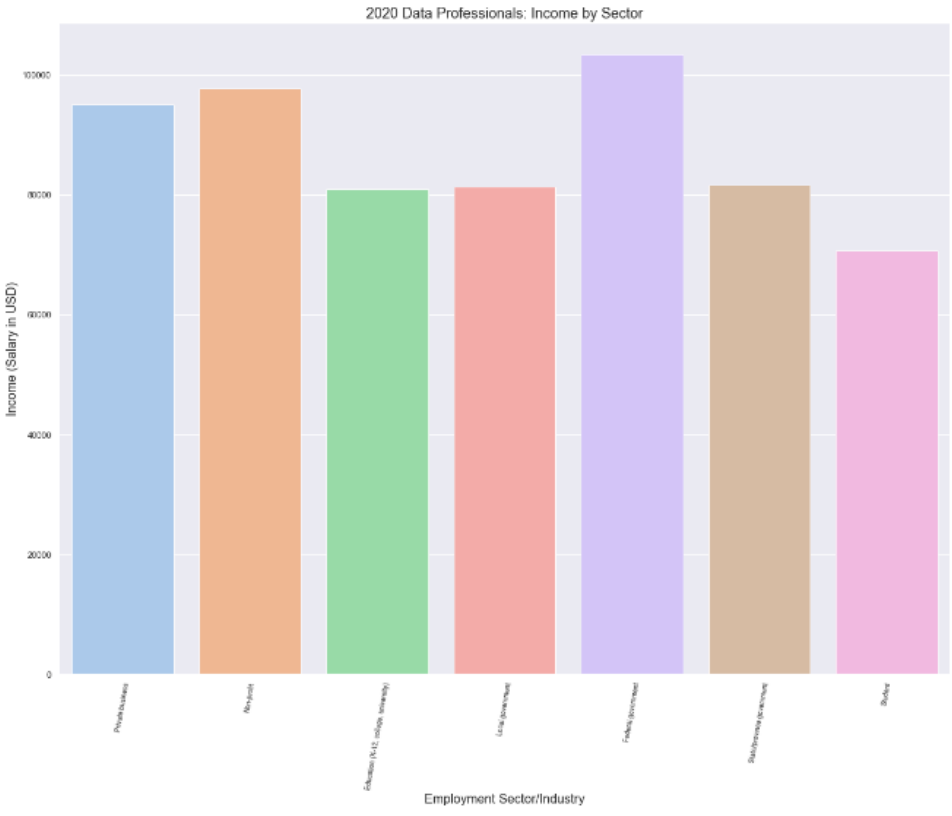

2020 Data Professionals: Boxplot of Income by Gender
1e6
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
SalaryUSD
Male
Female
Non-binary/third gender
Prefer not to say
NA
Not Asked
Gender

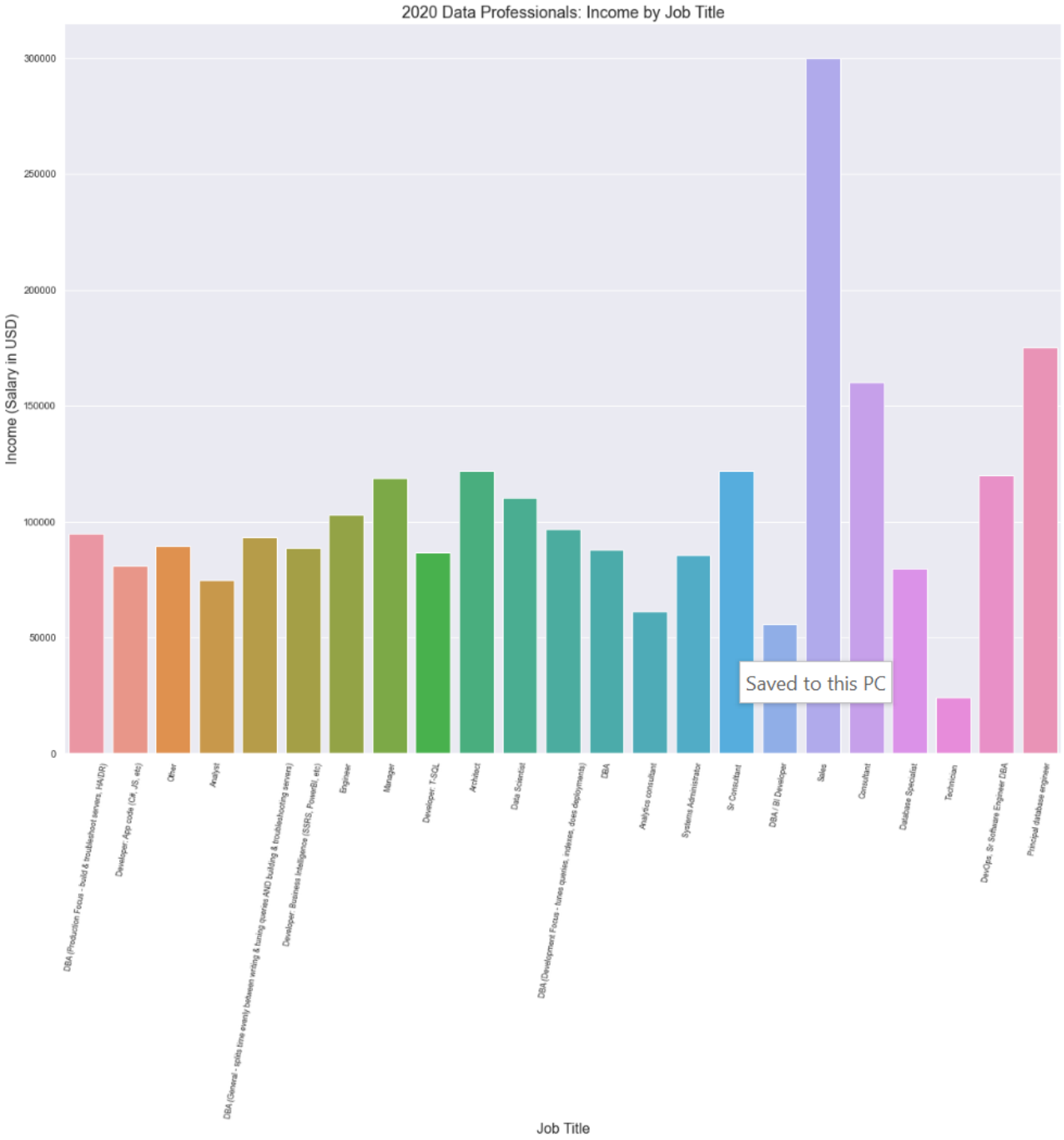

2020 Data Professionals: Distribution of Data Professionals by Country
Count
Country

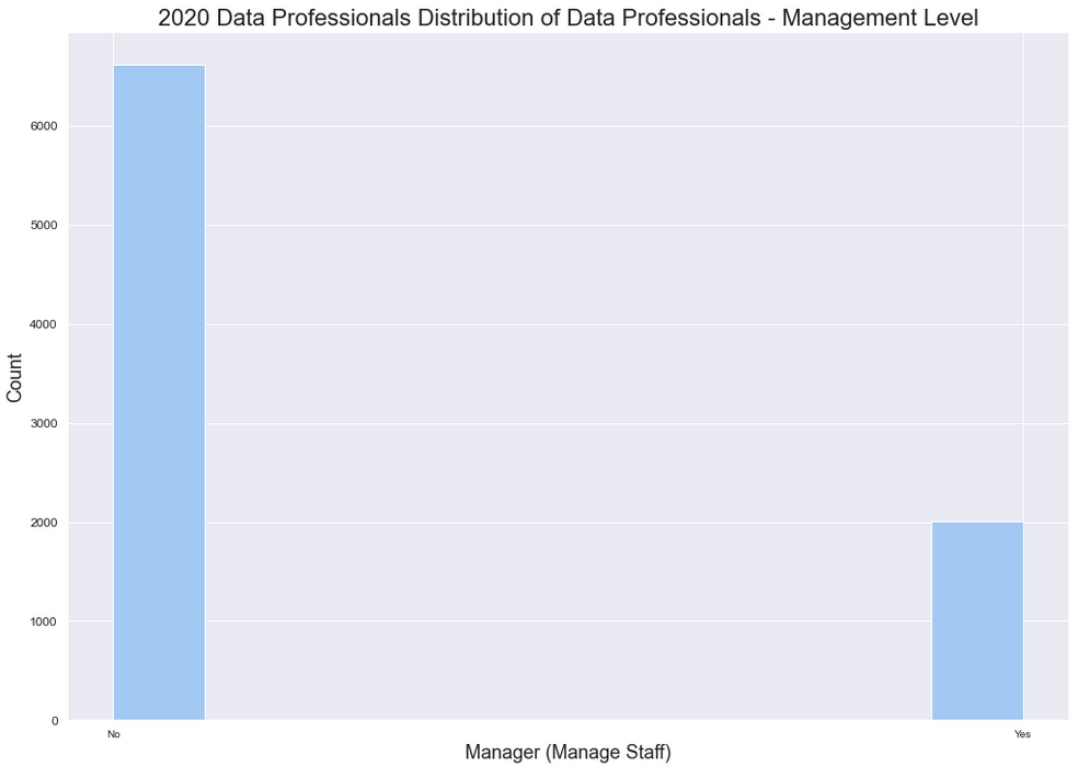

2020 Data Professionals: Income by Country
Income (Salary in USD)
Country

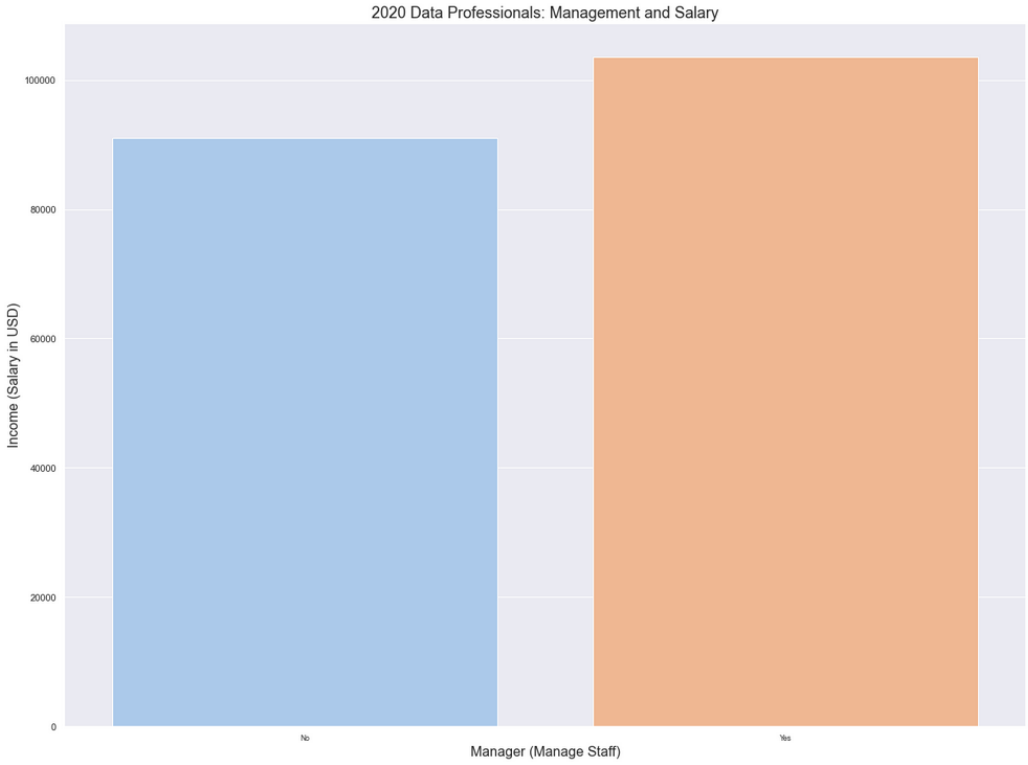

Distribution of Data Professionals by Job Title and Gender
Female
Male
NA
Non-binary/third gender
Not Asked
Prefer not to say
Count
0
200
400
600
800
1000
1200
1400
Analyst
Analytics consultant
Architect
Consultant
DBA
DBA (Development Focus - tunes queries, indexes, does deployments)
DBA (General - splits time evenly between writing & tuning queries AND building & troubleshooting servers)
DBA (Production Focus - build & troubleshoot servers, HA/DR)
DBA / BI Developer
Data Scientist
Database Specialist
DevOps, Sr Software Engineer DBA
Developer: App code (C#, JS, etc)
Developer: Business Intelligence (SSRS, PowerBI, etc)
Developer: T-SQL
Engineer
Manager
Other
Principal database engineer
Sales
Sr Consultant
Systems Administrator
Technician
Job Title

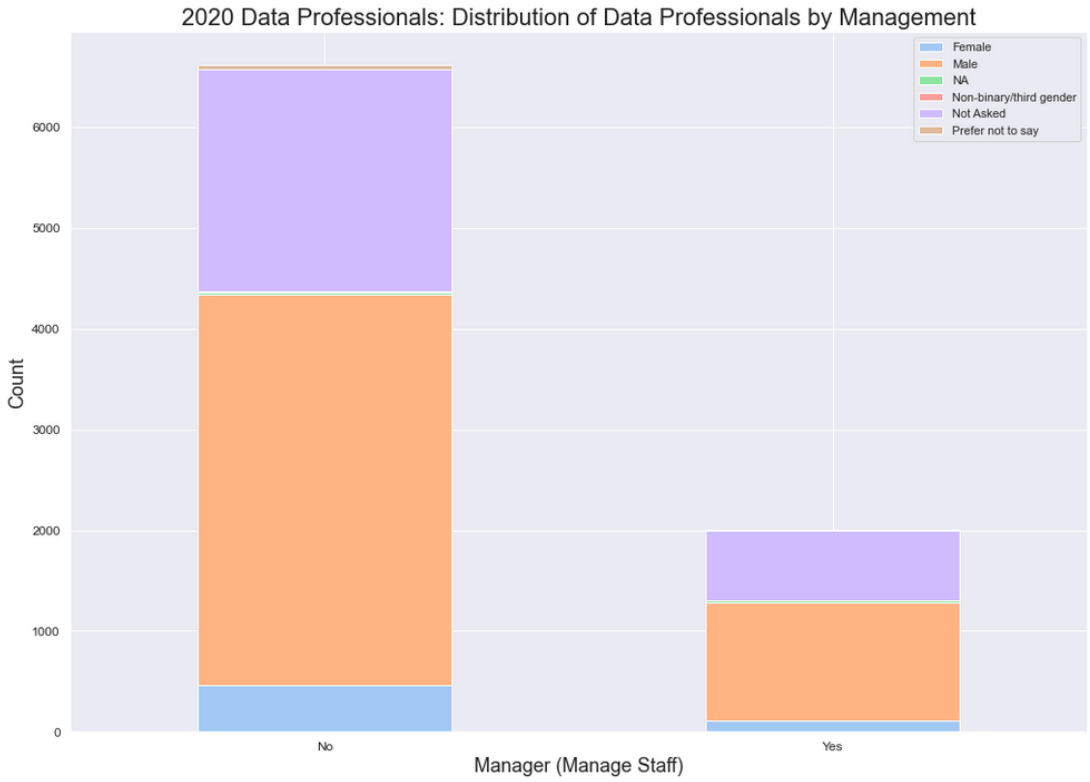

Distribution of Data Professionals by Sector and Gender
Female
Male
NA
Non-binary/third gender
Not Asked
Prefer not to say
Count
0
1000
2000
3000
4000
5000
6000
7000
Education (K-12, college, university)
Federal government
Local government
Non-profit
Private business
State/province government
Student
Employment Sector/Industry


2020 Data Professionals: Income by Sector
Income (Salary in USD)
0
20000
40000
60000
80000
100000
Private business
Non-profit
Education (K-12, college, university)
Local government
Federal government
State/province government
Student
Employment Sector/Industry

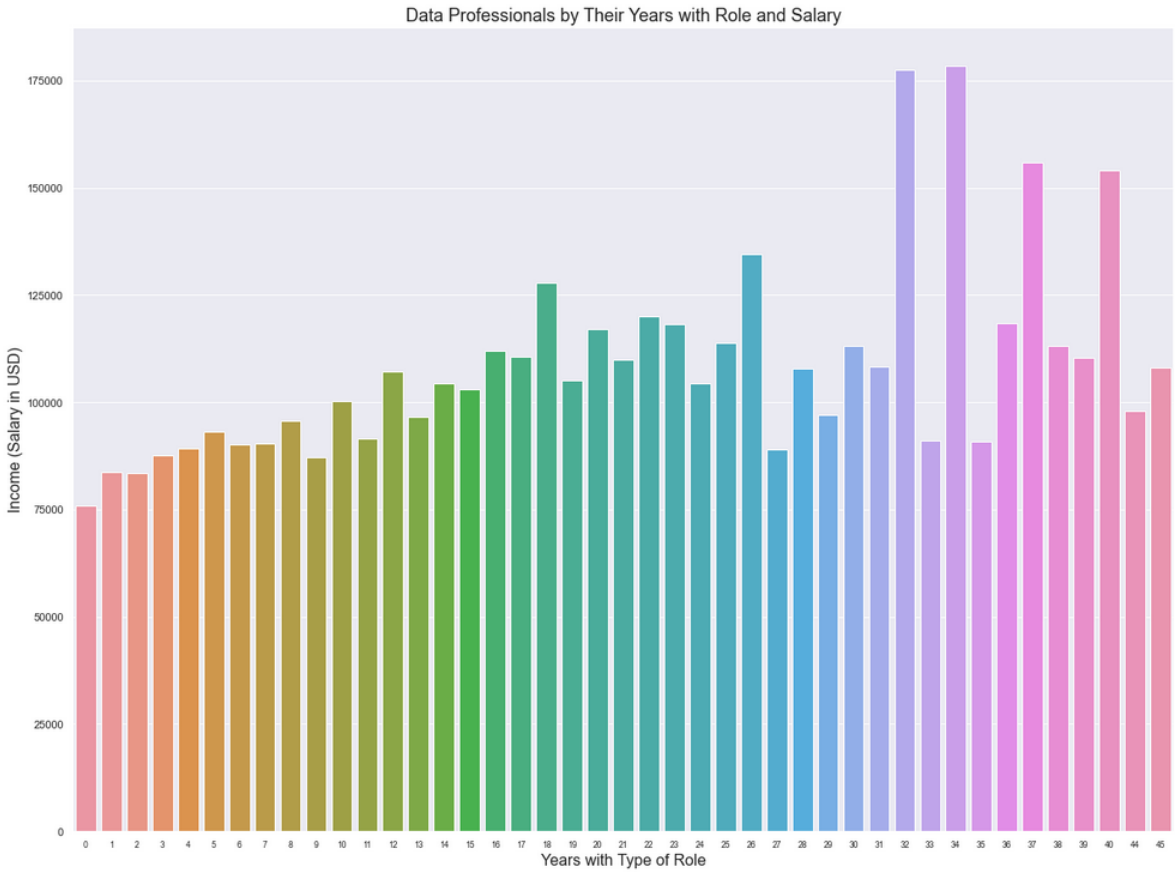

2020 Data Professionals: Income by Job Title
Income (Salary in USD)
300000
250000
200000
150000
100000
50000
0
Saved to this PC
DBA (Production Focus - build & troubleshoot servers, HA/DR)
Developer: App code (C#, JS, etc)
Other
Analyst
DBA (General - spids time evenly between writing & tuning queries AND building & troubleshooting servers)
Developer: Business Intelligence (SSRS, PowerBI, etc)
Engineer
Manager
Developer: T-SQL
Architect
Data Scientist
DBA (Development Focus - tunes queries, indexes, does deployments)
DBA
Analytics consultant
Systems Administrator
Sr Consultant
DBA / BI Developer
Sales
Consultant
Database Specialist
Technician
DevOps, Sr Software Engineer/DBA
Principal database engineer
Job Title

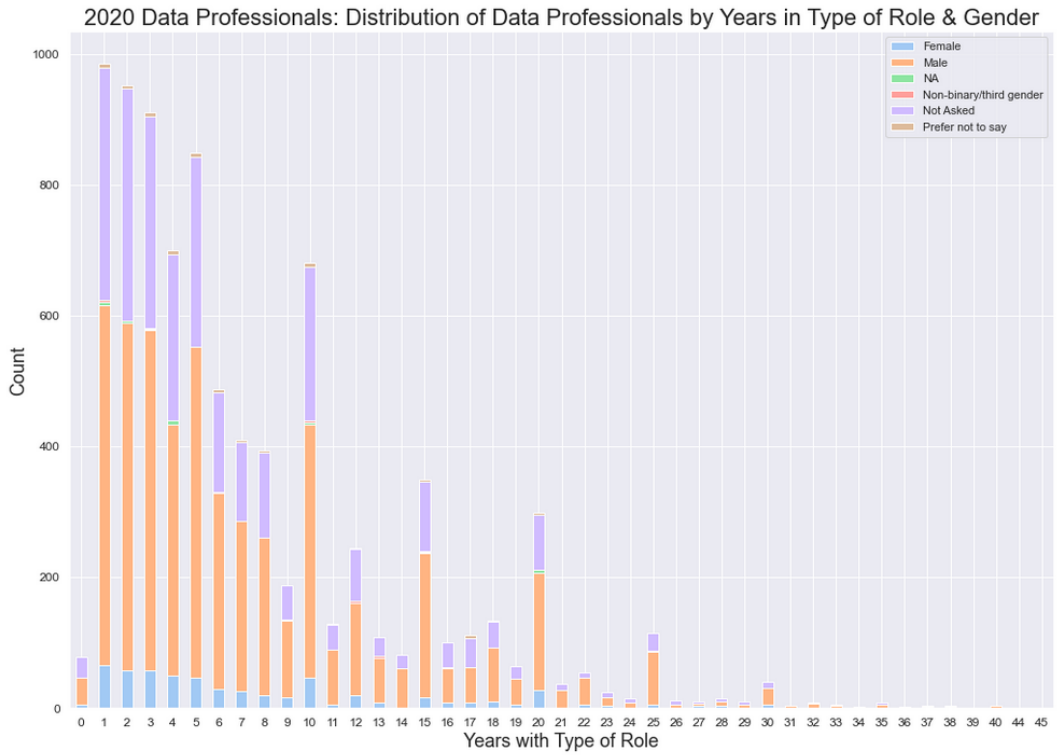

2020 Data Professionals Distribution of Data Professionals - Management Level
Count
6000
5000
4000
3000
2000
1000
0
No
Yes
Manager (Manage Staff)


2020 Data Professionals: Management and Salary
Income (Salary in USD)
100000
80000
60000
40000
20000
0
No
Yes
Manager (Manage Staff)


2020 Data Professionals: Distribution of Data Professionals by Management
Female
Male
NA
Non-binary/third gender
Not Asked
Prefer not to say
Count
0
1000
2000
3000
4000
5000
6000
No
Yes
Manager (Manage Staff)


Distribution of Data Professionals by Their Years with Role
Count
0
500
1000
1500
2000
2500
3000
3500
0
10
20
30
40
Years with Type of Role


Data Professionals by Their Years with Role and Salary
Income (Salary in USD)
Years with Type of Role


2020 Data Professionals: Distribution of Data Professionals by Years in Type of Role & Gender
Female
Male
NA
Non-binary/third gender
Not Asked
Prefer not to say
Count
Years with Type of Role

2. Descriptive Statistics Tables:

2020 U.S. Census Data:

	Number (people)	Median Income
count	8.000000	8.000000
mean	48724.250000	61903.125000
std	41705.495978	23167.212364
min	6963.000000	35574.000000
25%	20208.250000	45560.250000
50%	35134.000000	58259.000000
75%	67067.250000	72233.750000
max	129931.000000	101517.000000

	Number (people)	Median Income
count	4.000000	4.000000
mean	32007.500000	67751.500000
std	35921.752022	21769.964668
min	6987.000000	45870.000000
25%	14765.250000	52958.250000
50%	17853.500000	65116.500000
75%	35095.750000	79909.750000
max	85336.000000	94903.000000

Left Table Above: Gender and Household
Right Table Above: Race and Ethnicity

	Number (people)	Median Income
count	6.000000	6.000000
mean	21655.166667	69189.166667
std	9662.156228	18818.276567
min	5485.000000	46360.000000
25%	20906.250000	53056.000000
50%	21884.000000	72918.000000
75%	23778.250000	82838.000000
max	35688.000000	90359.000000

	Number (people)	Median Income
count	3.000000	3.000000
mean	38131.666667	72664.666667
std	9553.018389	30771.619917
min	31647.000000	47405.000000
25%	32646.500000	55529.000000
50%	33646.000000	63653.000000
75%	41374.000000	85294.500000
max	49102.000000	106936.000000

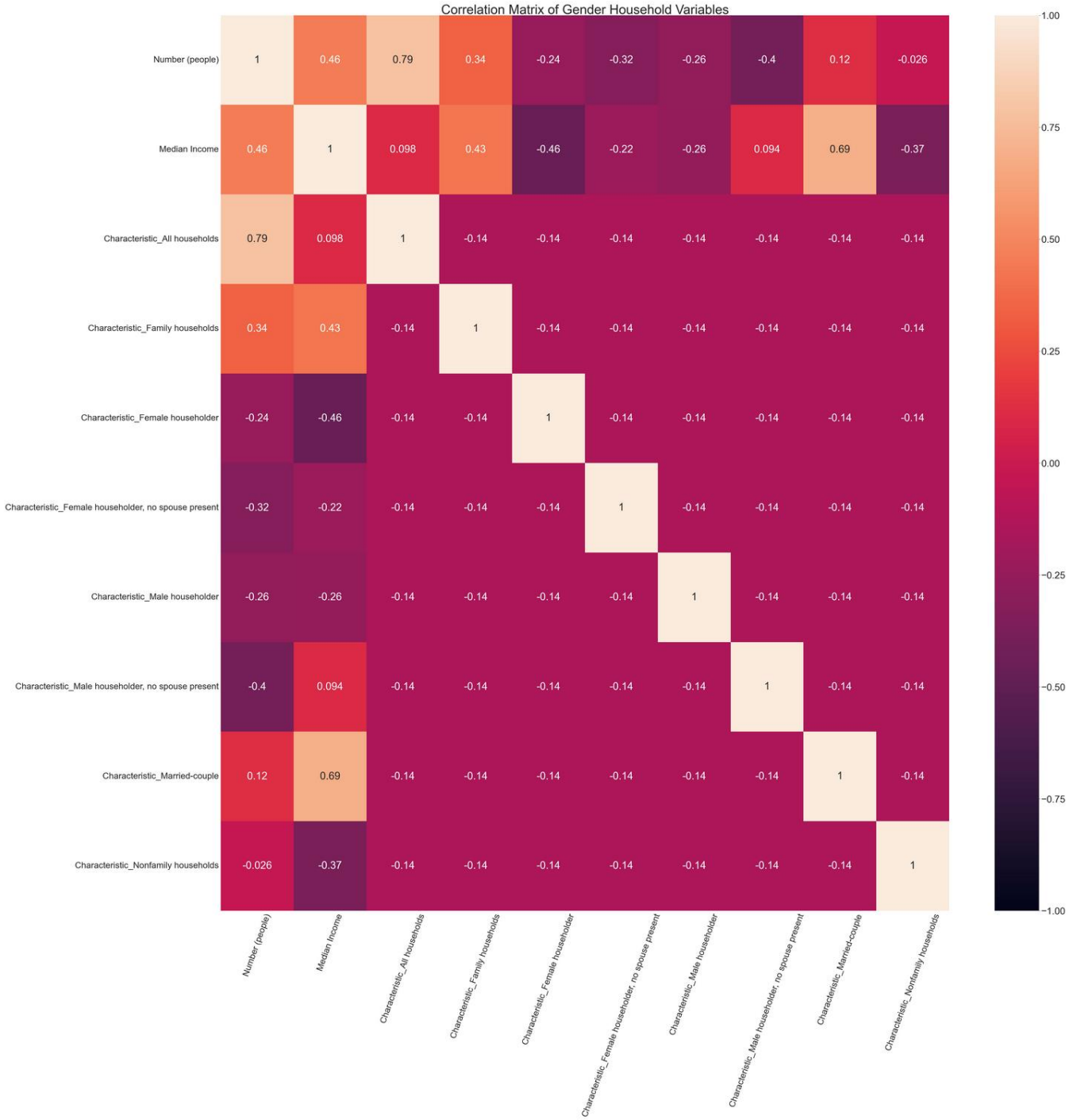
Left Table Above: Age
Right Table Above: Education

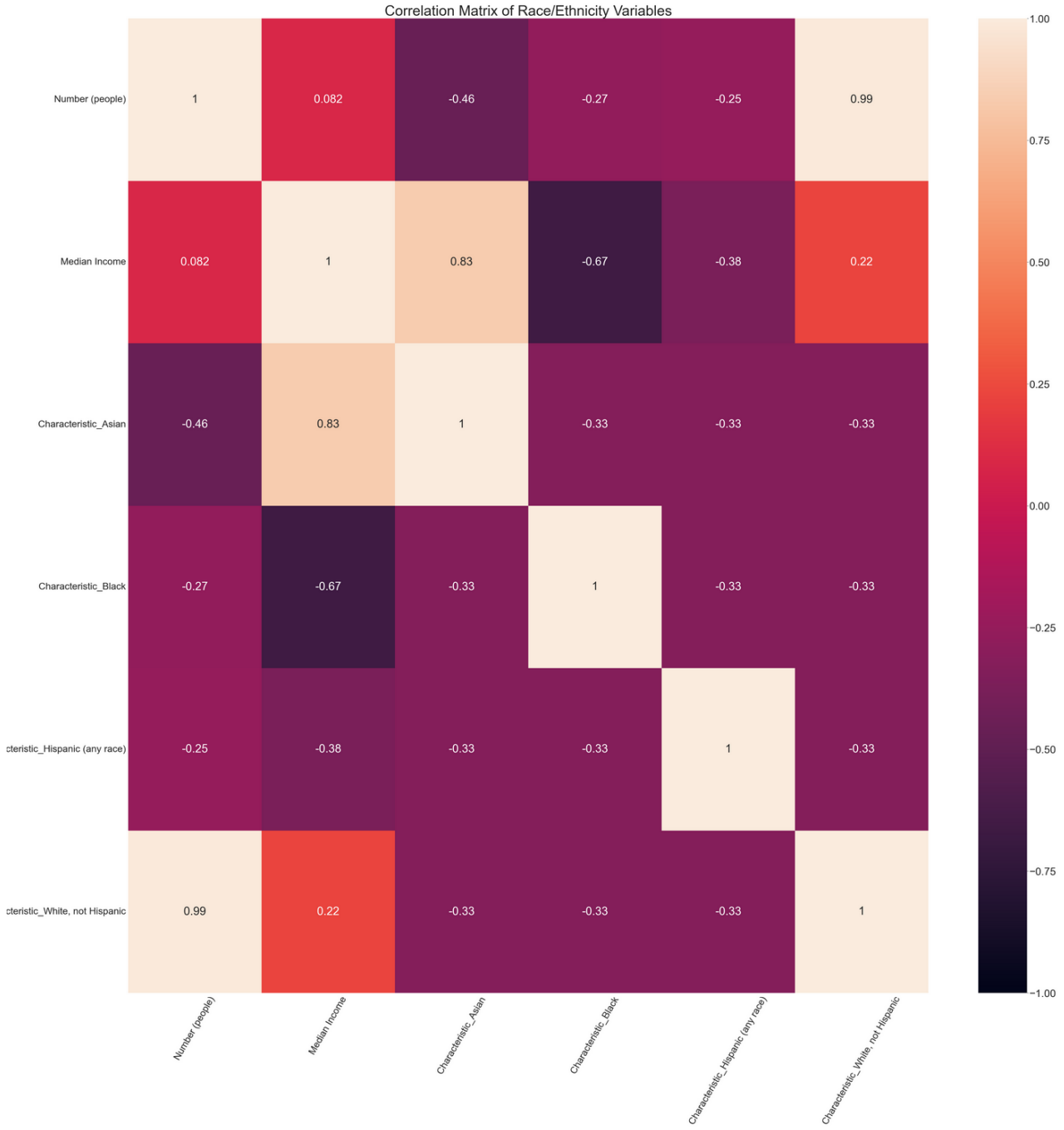
2020 Data Professionals Salary Survey Data:

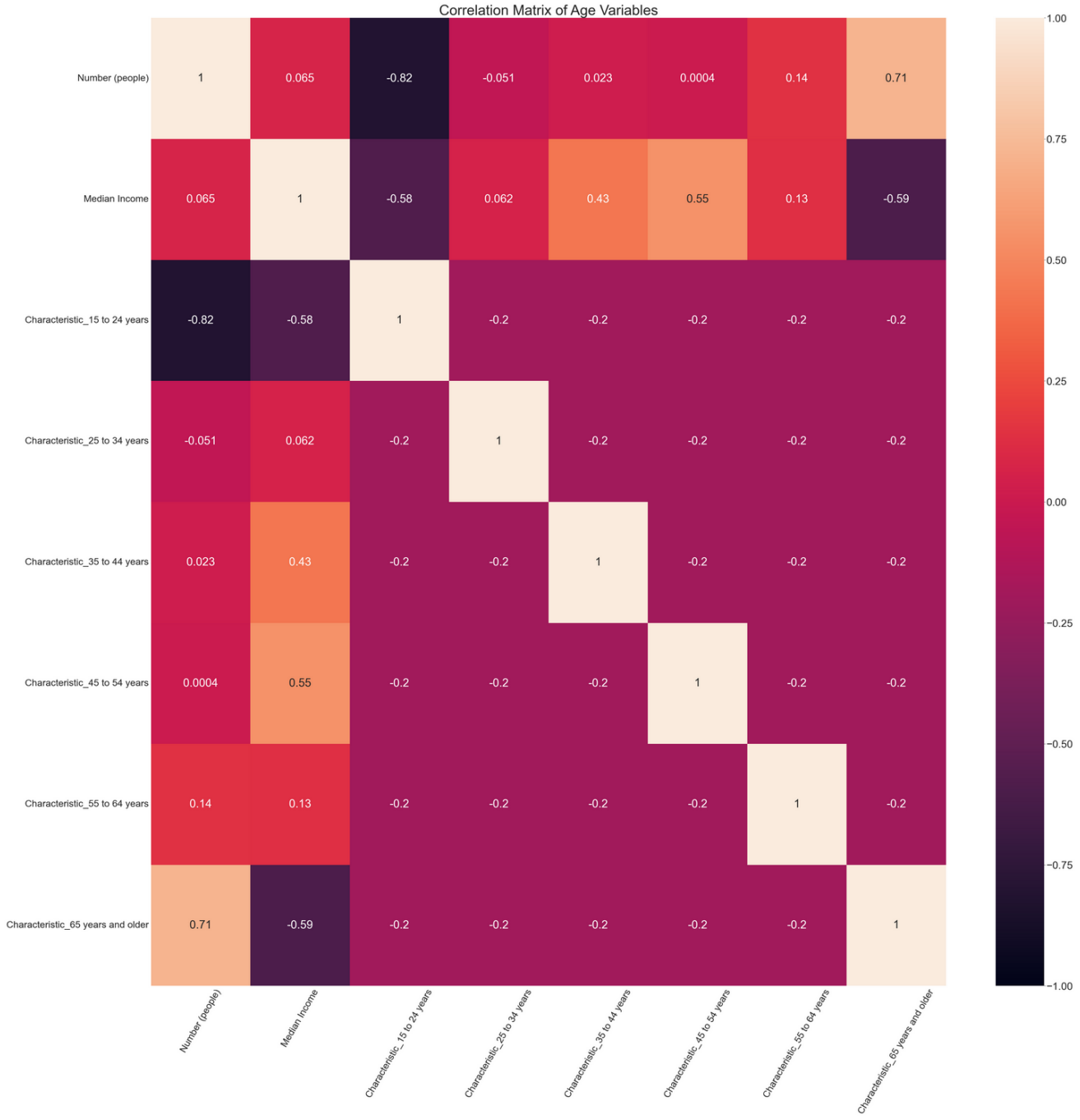
	SalaryUSD	YearsWithThisTypeOfJob
count	8.627000e+03	8627.000000
mean	9.394616e+04	7.544453
std	5.998446e+04	6.541186
min	0.000000e+00	0.000000
25%	6.500000e+04	3.000000
50%	9.000000e+04	5.000000
75%	1.150000e+05	10.000000
max	1.850000e+06	45.000000

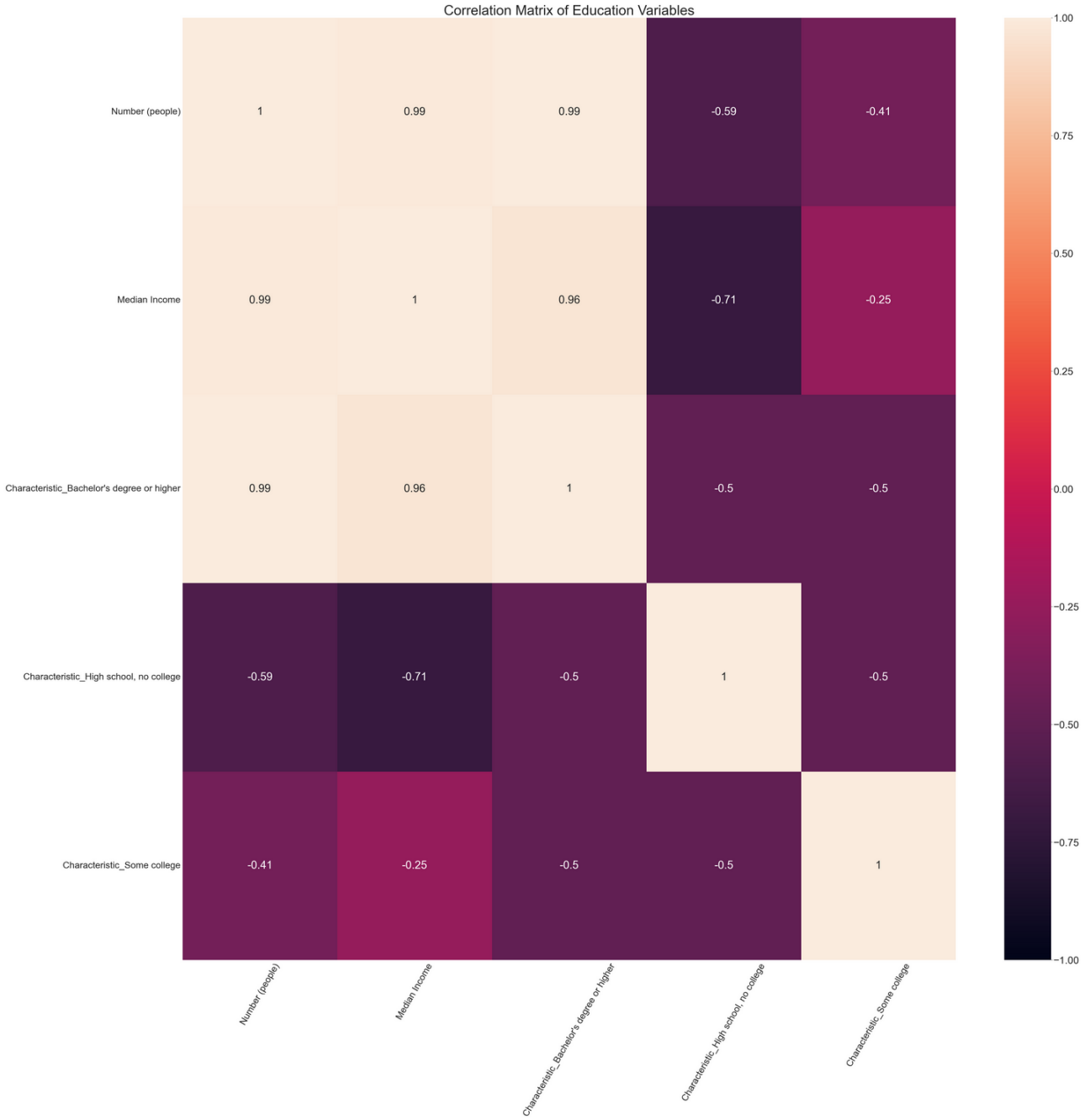
3. Correlation Heatmaps:

2020 U.S. Census Data:




Correlation Matrix of Race/Ethnicity Variables
Number (people)
Median Income
Characteristic_Asian
Characteristic_Black
cteristic_Hispanic (any race)
cteristic_White, not Hispanic
1 0.082 -0.46 -0.27 -0.25 0.99
0.082 1 0.83 -0.67 -0.38 0.22
-0.46 0.83 1 -0.33 -0.33 -0.33
-0.27 -0.67 -0.33 1 -0.33 -0.33
-0.25 -0.38 -0.33 -0.33 1 -0.33
0.99 0.22 -0.33 -0.33 -0.33 1
Number (people)
Median Income
Characteristic_Asian
Characteristic_Black
Characteristic_Hispanic (any race)
Characteristic_White, not Hispanic
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


Correlation Matrix of Age Variables
Number (people)
Median Income
Characteristic_15 to 24 years
Characteristic_25 to 34 years
Characteristic_35 to 44 years
Characteristic_45 to 54 years
Characteristic_55 to 64 years
Characteristic_65 years and older
1 0.065 -0.82 -0.051 0.023 0.0004 0.14 0.71
0.065 1 -0.58 0.062 0.43 0.55 0.13 -0.59
-0.82 -0.58 1 -0.2 -0.2 -0.2 -0.2 -0.2
-0.051 0.062 -0.2 1 -0.2 -0.2 -0.2 -0.2
0.023 0.43 -0.2 -0.2 1 -0.2 -0.2 -0.2
0.0004 0.55 -0.2 -0.2 -0.2 1 -0.2 -0.2
0.14 0.13 -0.2 -0.2 -0.2 -0.2 1 -0.2
0.71 -0.59 -0.2 -0.2 -0.2 -0.2 -0.2 1
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


Correlation Matrix of Education Variables
Number (people)
Median Income
Characteristic_Bachelor's degree or higher
Characteristic_High school, no college
Characteristic_Some college
1
0.99
0.99
-0.59
-0.41
0.99
1
0.96
-0.71
-0.25
0.99
0.96
1
-0.5
-0.5
-0.59
-0.71
-0.5
1
-0.5
-0.41
-0.25
-0.5
-0.5
1
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

4. Regression Outputs:

2020 Professionals Salary Survey Data:

OLS Regression Results

Dep. Variable:	SalaryUSD	R-squared:	0.002
Model:	OLS	Adj. R-squared:	0.001
Method:	Least Squares	F-statistic:	3.127
Date:	Mon, 11 Apr 2022	Prob (F-statistic):	0.00801
Time:	13:00:33	Log-Likelihood:	-1.0715e+05
No. Observations:	8627	AIC:	2.143e+05
Df Residuals:	8621	BIC:	2.143e+05
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	9.786e+04	2524.246	38.768	0.000	9.29e+04	1.03e+05
Gender[T.Male]	-2459.5973	2661.210	-0.924	0.355	-7676.205	2757.010
Gender[T.NA]	-781.4183	1.14e+04	-0.068	0.945	-2.32e+04	2.16e+04
Gender[T.Non-binary/third gender]	8621.6277	1.57e+04	0.550	0.583	-2.21e+04	3.94e+04
Gender[T.Not Asked]	-7144.7316	2758.964	-2.590	0.010	-1.26e+04	-1736.502
Gender[T.Prefer not to say]	-1.196e+04	8079.915	-1.480	0.139	-2.78e+04	3881.642

Omnibus:	13710.605	Durbin-Watson:	1.917
Prob(Omnibus):	0.000	Jarque-Bera (JB):	15251141.493
Skew:	10.027	Prob(JB):	0.00
Kurtosis:	208.002	Cond. No.	29.7

OLS Regression Results

Dep. Variable:	SalaryUSD	R-squared:	0.008
Model:	OLS	Adj. R-squared:	0.008
Method:	Least Squares	F-statistic:	68.40
Date:	Mon, 11 Apr 2022	Prob (F-statistic):	1.54e-16
Time:	13:00:33	Log-Likelihood:	-1.0712e+05
No. Observations:	8627	AIC:	2.142e+05
Df Residuals:	8625	BIC:	2.143e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	9.101e+04	734.545	123.905	0.000	8.96e+04	9.25e+04
ManageStaff[T.Yes]	1.259e+04	1521.774	8.270	0.000	9602.268	1.56e+04

Omnibus:	13734.350	Durbin-Watson:	1.914
Prob(Omnibus):	0.000	Jarque-Bera (JB):	15538784.384
Skew:	10.057	Prob(JB):	0.00
Kurtosis:	209.939	Cond. No.	2.52

Left Above: Gender + Salary
Right Above: Management + Salary

OLS Regression Results

Dep. Variable:	SalaryUSD	R-squared:	0.010
Model:	OLS	Adj. R-squared:	0.009
Method:	Least Squares	F-statistic:	14.24
Date:	Mon, 11 Apr 2022	Prob (F-statistic):	3.20e-16
Time:	13:00:33	Log-Likelihood:	-1.0711e+05
No. Observations:	8627	AIC:	2.142e+05
Df Residuals:	8620	BIC:	2.143e+05
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	9.55e+04	2530.162	37.743	0.000	9.05e+04	1e+05
ManageStaff[T.Yes]	1.271e+04	1521.959	8.349	0.000	9723.705	1.57e+04
Gender[T.Male]	-3059.7413	2651.642	-1.154	0.249	-8257.595	2138.112
Gender[T.NA]	-3673.8472	1.14e+04	-0.323	0.747	-2.6e+04	1.86e+04
Gender[T.Non-binary/third gender]	7598.7505	1.56e+04	0.486	0.627	-2.3e+04	3.82e+04
Gender[T.Not Asked]	-7804.5485	2749.171	-2.839	0.005	-1.32e+04	-2415.515
Gender[T.Prefer not to say]	-1.334e+04	8049.615	-1.657	0.097	-2.91e+04	2438.293

Omnibus:	13731.906	Durbin-Watson:	1.918
Prob(Omnibus):	0.000	Jarque-Bera (JB):	15516064.545
Skew:	10.054	Prob(JB):	0.00
Kurtosis:	209.787	Cond. No.	30.3

Left Above: Gender + Management + Salary

Appendix D:

References:

Daugherty, G. (2022, March 14). *Gender and income inequality*. Investopedia. https://www.investopedia.com/history-gender-wage-gap-america-5074898

Gould, E., Schieder, J., and Geier, K. (2016, October 20). *What is the gender pay gap and is it real?* Economic Policy Institute. https://www.epi.org/publication/what-is-the-gender-pay-gap-and-is-it-real/

Lyons, S. (2019, September 9). *The benefits of creating a diverse workforce.* Forbes. https://www.forbes.com/sites/forbescoachescouncil/2019/09/09/the-benefits-of-creating-a-diverse-workforce/?sh=2f65158f140b

Brent Ozar Unlimited. (2020). 2020 Data Professionals Salary Survey Data. https://www.brentozar.com/archive/2020/01/the-2020-data-professional-salary-survey-results-are-in/

U.S. Census Bureau. (2020). Table 1A: Educational Attainment, People 18 Years Old and Over by Total Money Earnings in 2020 by Work Experience, Age, Race, Hispanic Origin, and Sex. https://www.census.gov/data/tables/2020/demo/educational-attainment/cps-detailed-tables.html

Appendix E:

Audience Questions:

1. Did any relationships exist between any of the variables outside of the wage/income variable? What variables had the highest multi-collinearity? How might this impact any outcomes?
    a. From my analyses, there were certainly correlations between population and aggregate income amounts (median). However, beyond that, there did not appear to be any significant relationships between the independent variables themselves.
    b. As for multi-collinearity, this was handled via the use of dummy variables, so did not end up presenting a concern.
    c. I think the greatest impact to any outcomes was the imbalanced variables classes and any potential skew from that. In the future, using a SMOTE methodology with ML techniques might help here.

2. How did you determine which variables were most important prior to your analysis? Do you feel there were any you missed that might have been important to include?
    a. My determination for the most important variables primarily came from the following:
        i. My literature review of what was most important to explore.
        ii. The variables available to me in the data.
        iii. Those variables that I ascertained might impact income.
    b. Instead of merely just missed variables, I feel the greatest lack within my project is the amount of and type of data I had to work with. I feel I could have had more robust outcomes with some better data.
3. What was your initial hypothesis, and do you feel the results and outcome of your project work supports this? Why or why not?
    a. Initially, I did hypothesize that gender and race would most certainly still present a prevailing income gap in 2020. I did not necessarily know how this might look specifically.
    b. I also hypothesized that other variables beyond race and gender would impact the income one earns.
    c. I do feel as though a portion of my results (primarily the U.S. Census data results), because those results did show evidence to support my hypotheses.
4. What do you recommend for those who want to help in this area? What other ways can those who are interested use data to inform others and instill change?
    a. Firstly, I recommend more research, and continued research.
    b. I also recommend that research findings be taken into serious consideration – research without action never amounts to much in the real-world.
5. Talk a bit more about your regression and/or ML model methods – how did you ensure that the data was transformed appropriately and then properly fed into those models?
    a. For my regression analyses, I had to ensure that my data was properly cleaned but also that I used dummy variables in order to control for any multi-collinearity prior to the analysis portion.
6. Why do you think this topic is important to look into further than the research that already exists out there?
    a. I think more individual data would be useful – from a variety of sectors, companies, industries, etc. It can be tough to get the full picture without enough varied data. The data I used only gave two glimpses – into the U.S. in aggregate form, and into the income lives of data professionals globally. I think we need to continually expand upon just those two snapshots.
7. Do you believe that with additional data and additional research on this topic in coming years that it will continue to help push and drive change in this area? Why or why not?
    a. I do – especially when coupled with the socio-political movements that are currently taking place. Equity movements and missions are at the forefront – research is the motor and the fuel behind that.
8. Is there any other data that you wish was available that you couldn't find, or that was not in a usable format for you?

    a. Yes, I do wish that I was able to move forward with my original project plans using the PSID data. That data was much more individualized, but a lot harder to work with. I hope I can expand upon this project more in the future by including that data.
9. What challenges and limitations did you personally face during this project? In addition to that, what do you think are the overall challenges and limitations of your project work itself, and what do you think could be done next to improve upon that?
    a. It was a challenge to find good data with high usability.
    b. It was also a challenge to really find meaning from this data – of course, there were my findings, but I was left with a sense of feeling like I wanted more out of this. I think this is partially due to the kind of data I analyzed – I think I was merely curious about other data out there and what that might look like comparatively.
    c. Again, as mentioned previously, more analyses with additional data sources would be the best next step, I believe.
10. Why do you believe the results are what they are? What are resources or references either support or do not support your findings? Why do you think your results are what they are?
    a. For the U.S. Census data, I believe those results are merely a reflection of U.S. society and how we are still dealing with the lingering effects of unequal systems and systems of oppression. Our country was essentially founded upon these "imbalances", and we are still playing catch up.
    b. For the Data Professionals data – I think this data wasn't as robust and reliable as I would like, and it was quite imbalanced, so I felt the results of this to be somewhat inconclusive based upon that.