Deep Sea Asset Management Case Study:

Using Machine Learning to Better Target Wealth Prospects

Nicole Roberts, November 2024

Executive Summary: Insights for Deep Sea Asset Management

BUSINESS APPLICATIONS

1. **Targeted Marketing for Financial Products**:
    A. **Age and Working Status**: Knowing that people in the working age bracket with full-time jobs have higher income potential, the company can focus its marketing efforts on working professionals between the ages of 30 and 55. This demographic is more likely to have disposable income and be interested in financial services.
    B. **Education and Occupation**: Since higher education and executive roles are associated with higher income, the company could target individuals with advanced degrees or specific occupations (e.g., executives, managers) for its high-end advisory services. This can be done by targeting ads on professional networks like LinkedIn, especially among users with these qualifications.

2. **Product Development and Pricing Strategy**:
    A. **Capital Gains and Stock Dividends**: The insight that reporting capital gains and stock dividends is associated with higher income could lead the company to develop specialized investment products or portfolios aimed at high-net-worth individuals. These products could focus on tax-efficient investments, targeting wealth clients who report stock dividends or capital gains to help them manage and grow their wealth more effectively.
    B. **Head of Household**: The company could create packages or advisory services tailored specifically for heads of households, who may be making more financial decisions for their families. These packages might include estate and legacy planning, education savings plans, and life insurance to secure their family's future.

3. **Client Retention Strategies**:
    A. **Gender**: Understanding that men might generally earn more could shape how the company approaches retention strategies. They might create specialized programs or incentives for male clients in their 30s and 40s who have been identified as high earners. Simultaneously, they could createtargeted outreach programs to support female clients in reaching their financial goals, as an opportunity to promote gender-specific financial planning services.
    B. **Having Children at Home**: The firm could design retention strategies focusing on clients with children at home by offering college savings plans, educational workshops on family financial planning, or loyalty programs that reward long-term clients who invest in family-oriented financial products.

4. **Choosing Where to Expand Services**:
    A. **Very Large Companies**: Since income is associated with working for large companies, the wealth management firm might consider partnerships with big corporations to attract clients who have more disposable income to invest. They could offer on-site or virtual seminars on financial wellness, investment options, and retirement planning for employees of these companies.

5. **Tax-Optimized Financial Advice**:
    A. **Married and Filing Joint Taxes**: The finding about marriage and filing status could prompt the company to offer more specialized tax-optimized financial advice for married couples. They might develop tax-efficient investment portfolios or consultative services to help couples maximize their joint income, save on taxes, and achieve financial goals together.

Agenda

Background, Purpose, & Approach

Background: Marketing Research to Target Wealthy Clientele

PROBLEM STATEMENT

Deep Sea Asset Management is a wealth management firm offering advisory services, financial planning, and retirement planning for individuals.

Their primary goal was to effectively target and maximize acquisition of higher-value* individuals, however they were unsure of how to so do. Specifically, they did not know where nor how their marketing efforts are best utilized in order to reach and resonate with higher-income segments.

SOLUTION

Deep Sea conducted **market research, including developing a random forest machine learning model,** to identify and help tailor services, products, and marketing strategies to align with the financial behaviors and needs of higher-income segments.

This market research aimed to derive insights as to **which demographic characteristics indicate higher value individuals**. Knowing which characteristics are the key features that predict clients are above a certain income threshold will **help them formulate effective marketing campaigns that are most likely to resonate with this target segment.**

**"Higher value" here does not indicate high-net worth.*

Background: Market Research Dataset, *The US Census*

The United States Census Bureau leads the country's Federal Statistical System; its primary responsibility is to collect demographic and economic data every 10 years about America to help inform strategic initiatives.

Market research conducted for Deep Sea Asset Management examined the characteristics of the Census's subpopulations across the country. This enabled insights into household income in relation to various demographic traits, including **which demographic traits *predict* income bracket**.

For Deep Sea, the target income bracket they would like to reach is **above $100K**.

ANALYSIS PURPOSE

To identify characteristics in US Census archive data that predict whether a person makes more or less than $100,000 per year. **Those who make more than $100K are the segment of interest for Deep Sea Asset Management.**

ANALYSIS APPROACH

Data was **explored**, **cleaned**, and **prepped** for a binary classification random forest model (to predict one of two classes: more than $100K / less than $100K). Initial **heuristics** were conducted followed by analysis with the random forest model.

Insights and conclusions were derived, including how the findings can be applied to **help Deep Sea Asset Management better target valuable customers through their marketing efforts**.

Data Deep Dive: Exploratory Analysis and Preparation

US Census: A stratified dataset

Stratification involves dividing the population into distinct subgroups, or **strata**, before sampling. Each stratum is typically homogeneous with respect to certain characteristics, but distinct from other strata.

US Census data is stratified, meaning **each datapoint represents a subpopulation**.

To reflect the *entire* population, and not just the sampled individuals, **this analysis incorporated the US Census dataset's *instance weight* field throughout (except where noted with '*strata*').**

The Raw Dataset

- **~293K distinct strata** total (once duplicates were removed).
- **42 demographic fields.**
- No missing data.

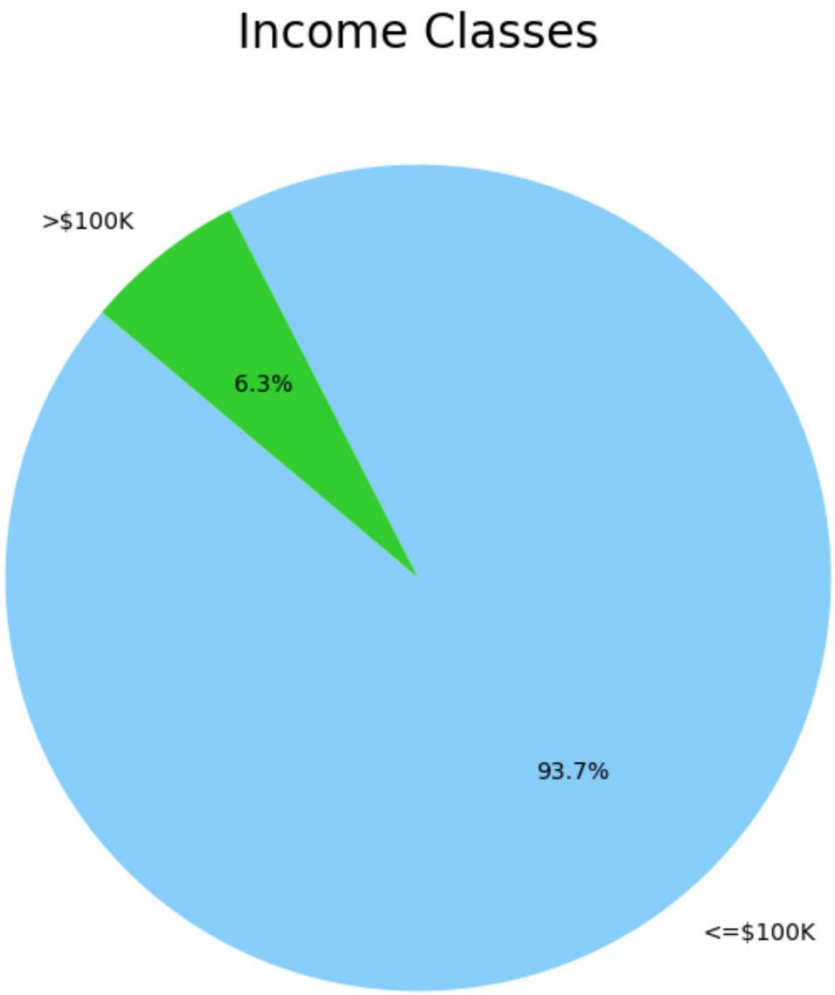
Data Prep

- Duplicates removed.
- Header names were defined.
- Column name typos, inconsistent across train and test datasets, were corrected.
- When possible, numericized categorical data was converted to text for ease of interpretation (e.g. ‘0’ indicates ‘Non-veteran’ for *veterans benefits).*
- Where applicable, data was cast into correct data types.

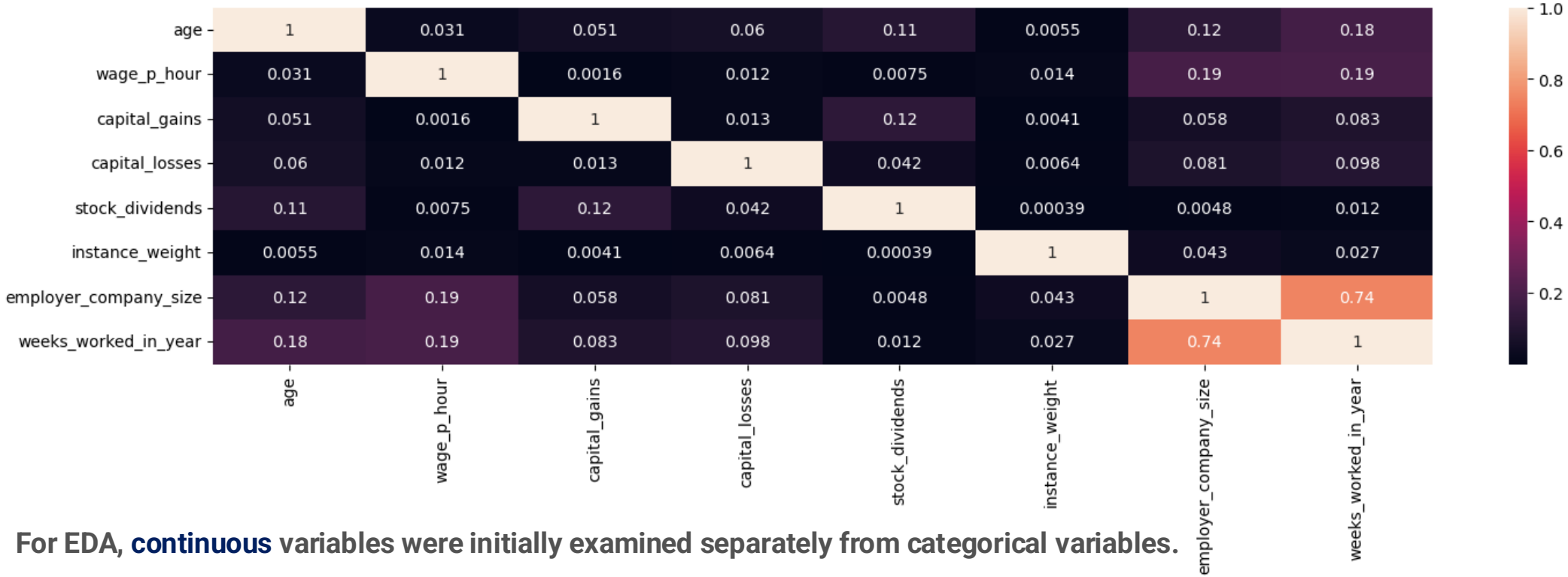
Data Deep Dive: Target Variable

Target Variable:
Income <=$100K, >$100K

- **Binary classification** problem
- Significant **class imbalance**, presents a challenge for machine learning applications:
  - Majority class dominance
  - Poor minority class performance

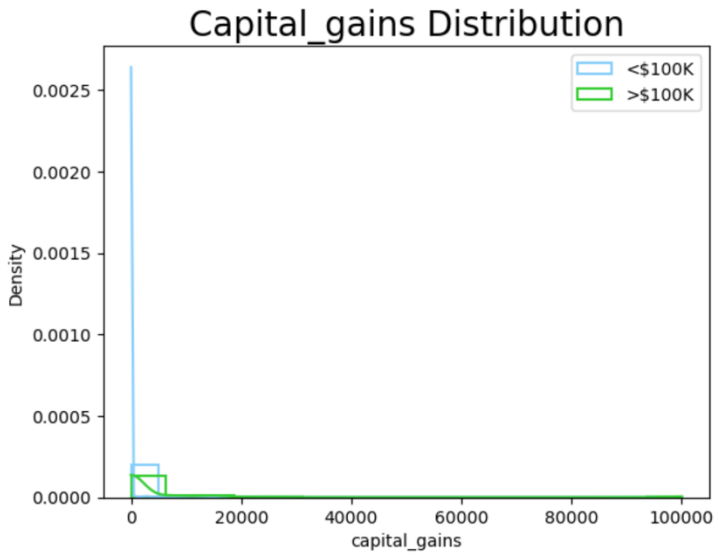


Data Deep Dive: Continuous Variables, Pearson Correlation Coefficients

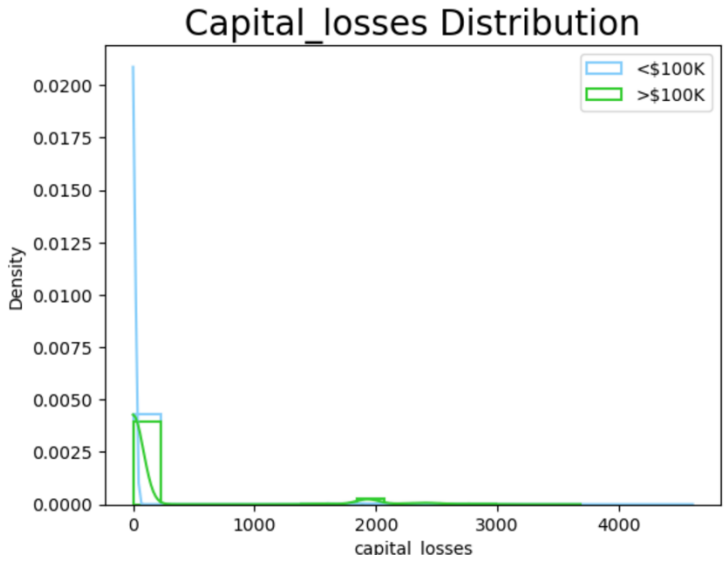


- **For EDA, continuous variables were initially examined separately from categorical variables.**
  - Note, *all* variables (both continuous and nominal) were analyzed in the context of their ability to effectively serve as features in machine learning models in the interest of eliminating those who are likely to add noise.
- **Pearson correlation coefficients** indicate negligible-to-weak relationships between continuous features – with the exception of *employer company size* and weeks *worked in year*.
  - **Multicollinearity** can negatively impact certain machine learning models, e.g. Logistic Regression.

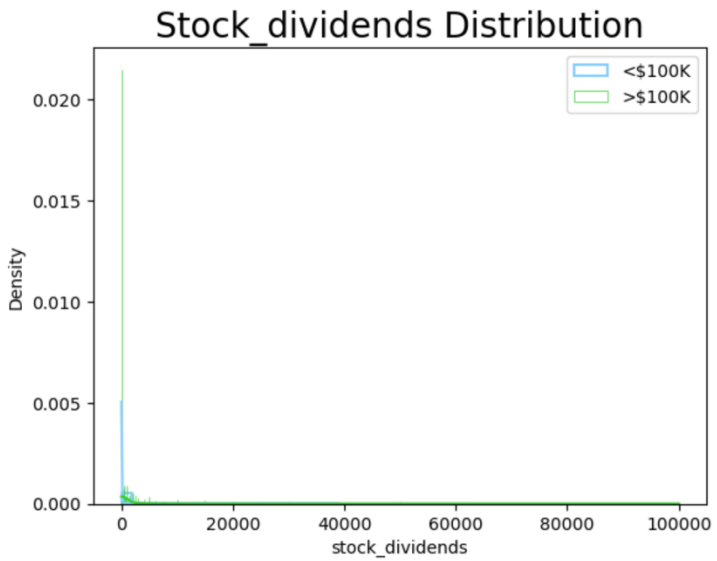
Data Deep Dive: Distributions of Continuous Variables, Capital Gains/Losses, & Stock Dividends



Point Biserial Correlation with Income: **0.24**



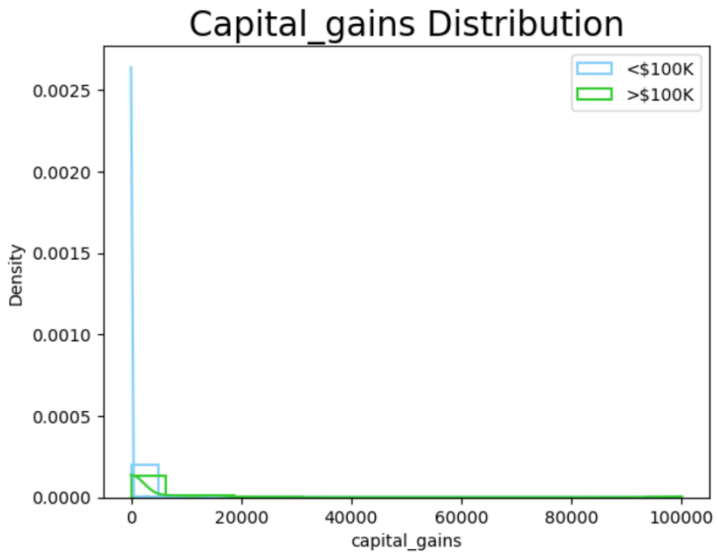
Point Biserial Correlation with Income: **0.15**



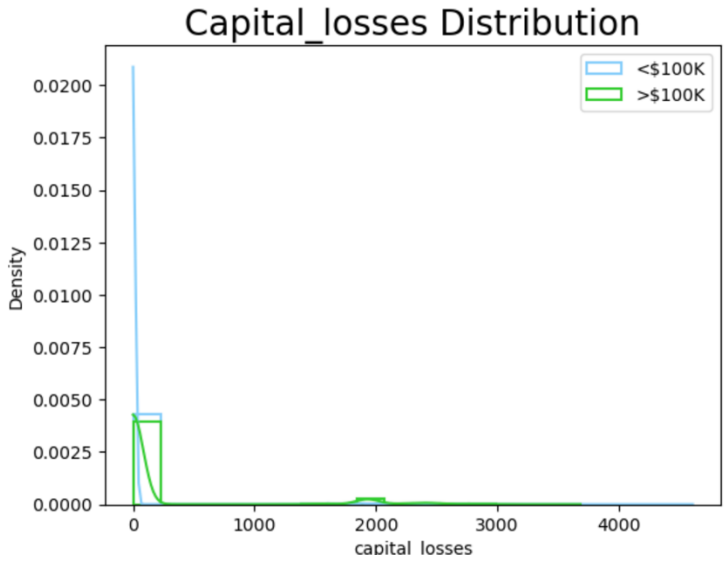
Point Biserial Correlation with Income: **0.18**

- The percent of strata who have **no capital gains** to report are **96.2%.**
- The percent of strata who have **no capital losses** to report are **98.0%.**
- The percent of strata who have **no stock dividends** to report are **89.2%.**

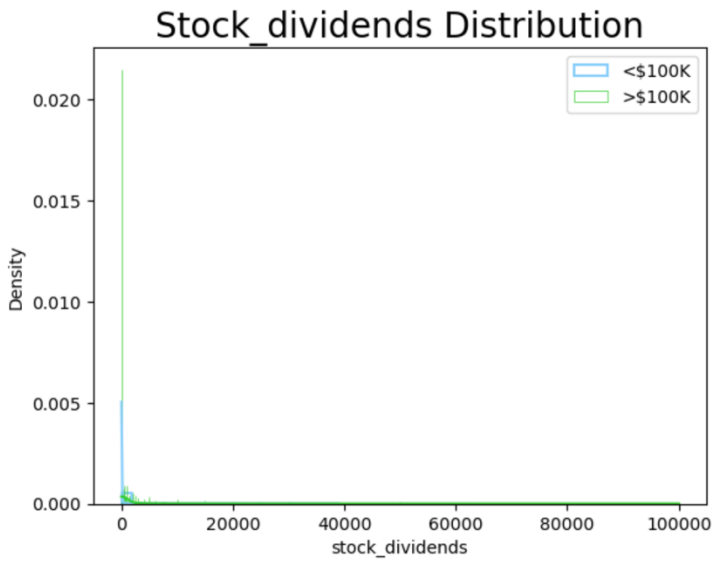
Data Deep Dive: Distributions of Continuous Variables, Capital Gains/Losses, & Stock Dividends



Point Biserial Correlation with Income: **0.24**



Point Biserial Correlation with Income: **0.15**



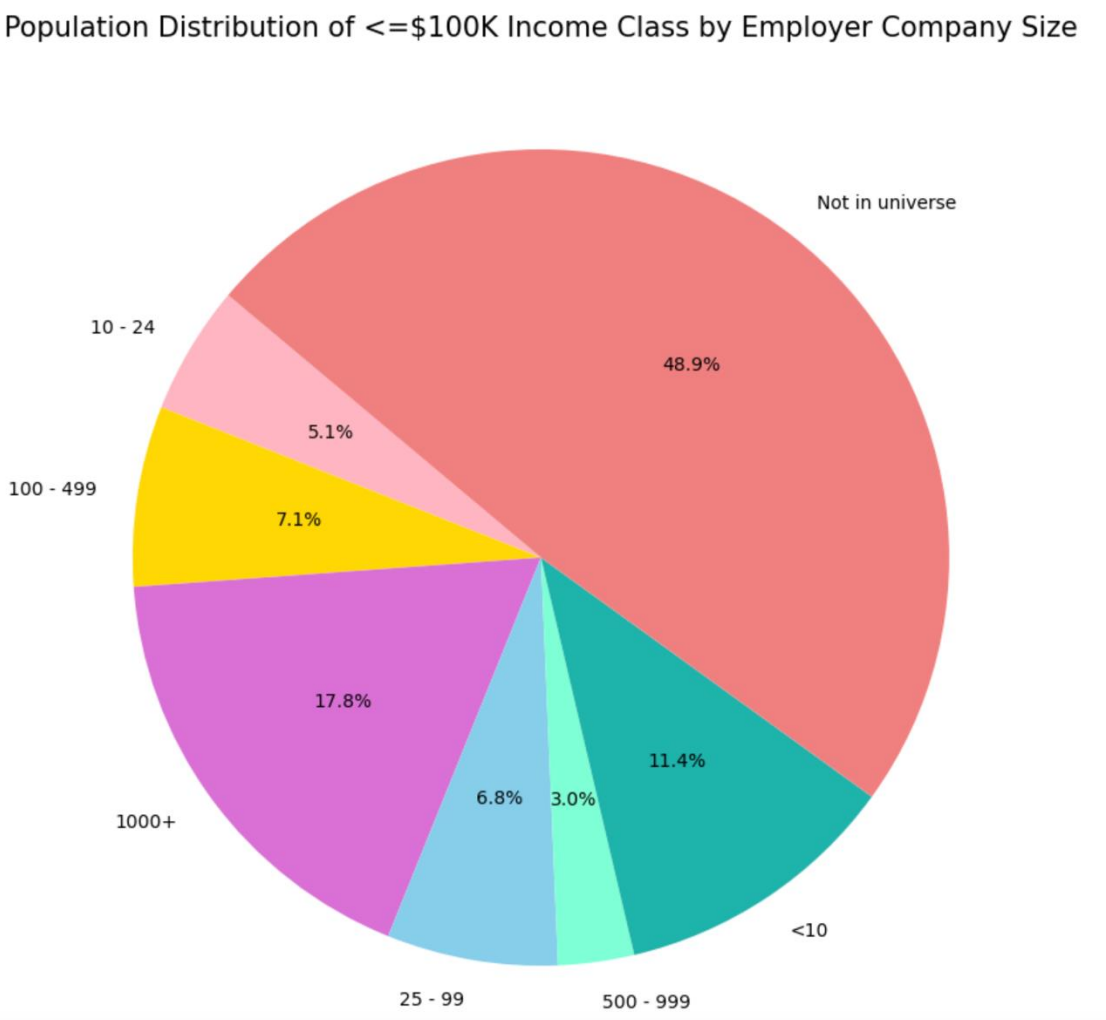
Point Biserial Correlation with Income: **0.18**

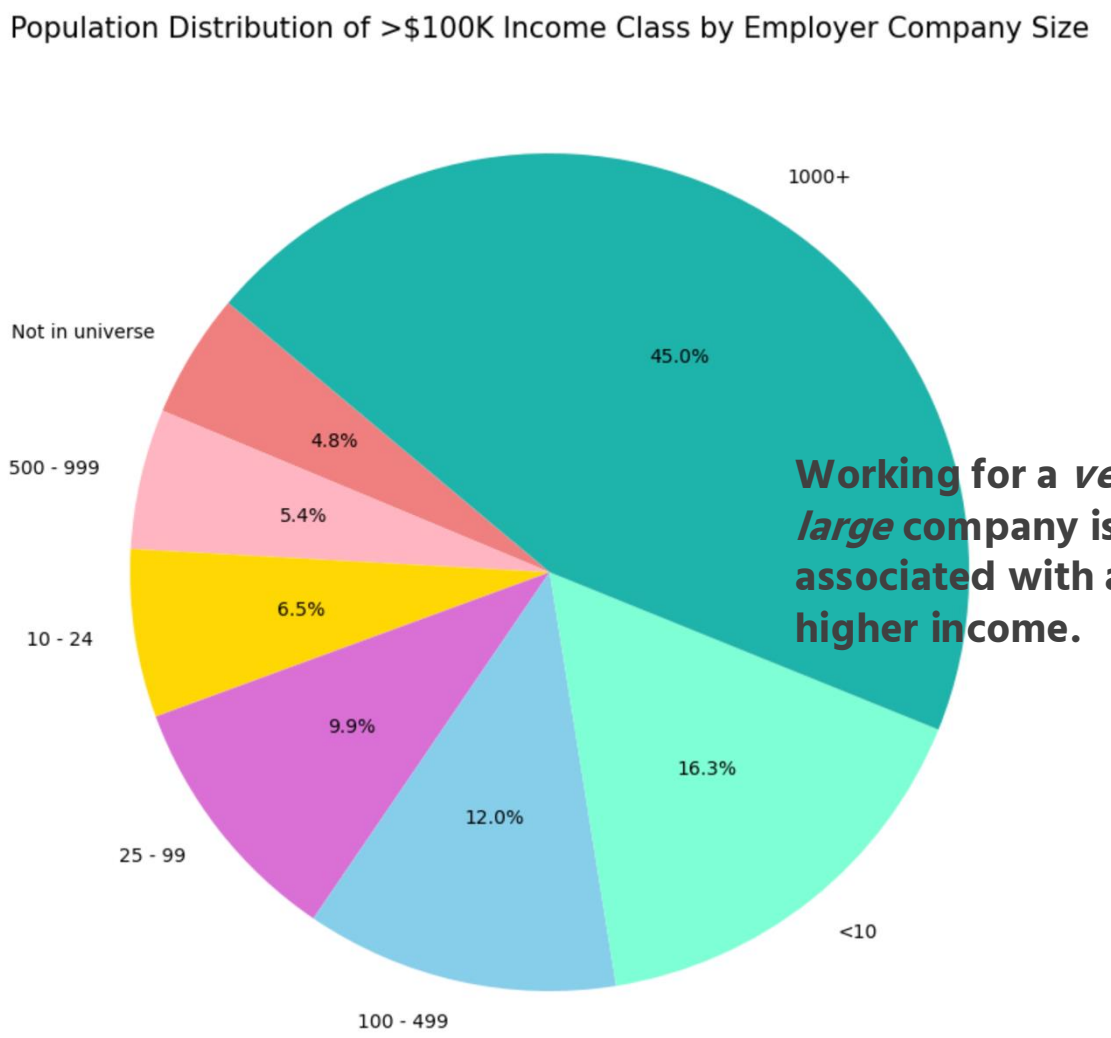
- The percent of strata who have **no capital gains** to report are **96.2%.**
- The percent of strata who have **no capital losses** to report are **98.0%.**
- The percent of strata who have **no stock dividends** to report are **89.2%.**

Converting from continuous to binary categories *could* improve model performance.

Data Deep Dive: Employer Company Size Population Distribution

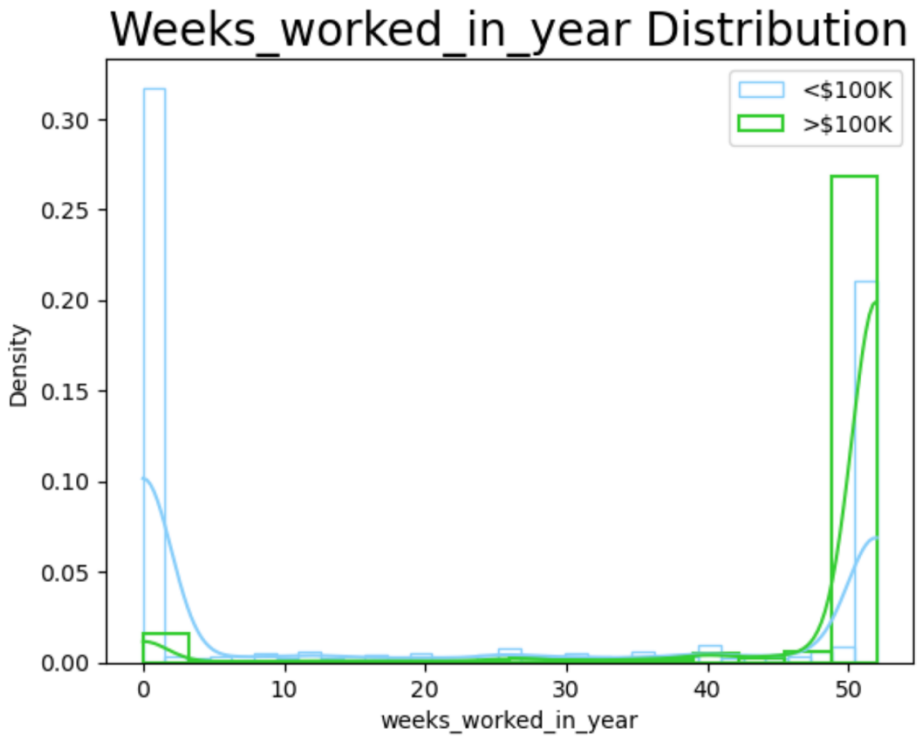
** Employer company size was converted into a categorical variable.*





Working for a *very large* company is associated with a higher income.

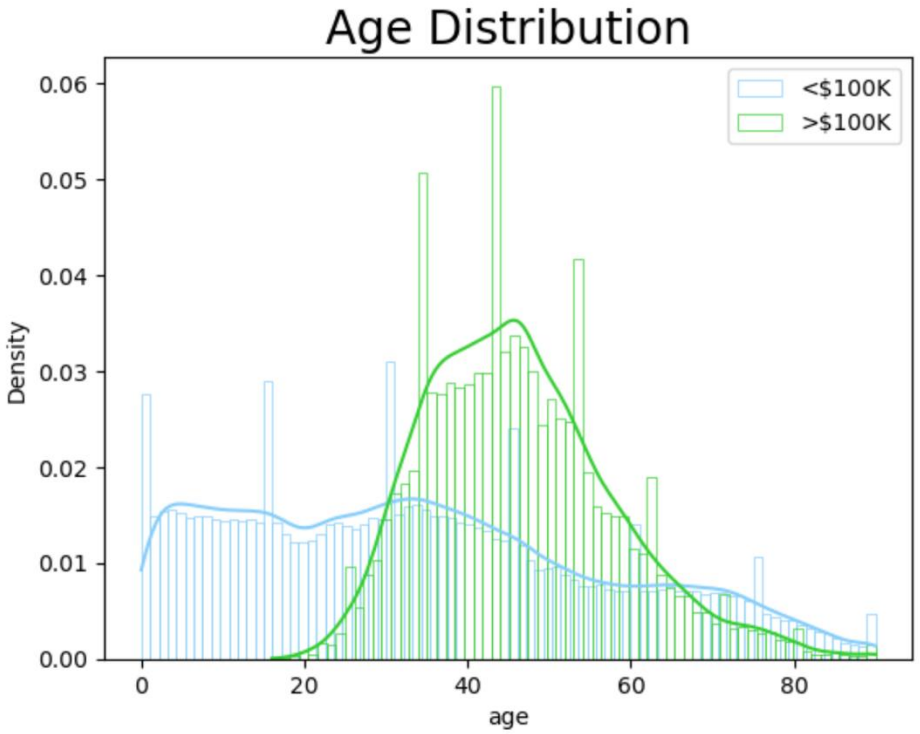
Data Deep Dive: Distributions of Continuous Variables, Weeks Worked p. Year

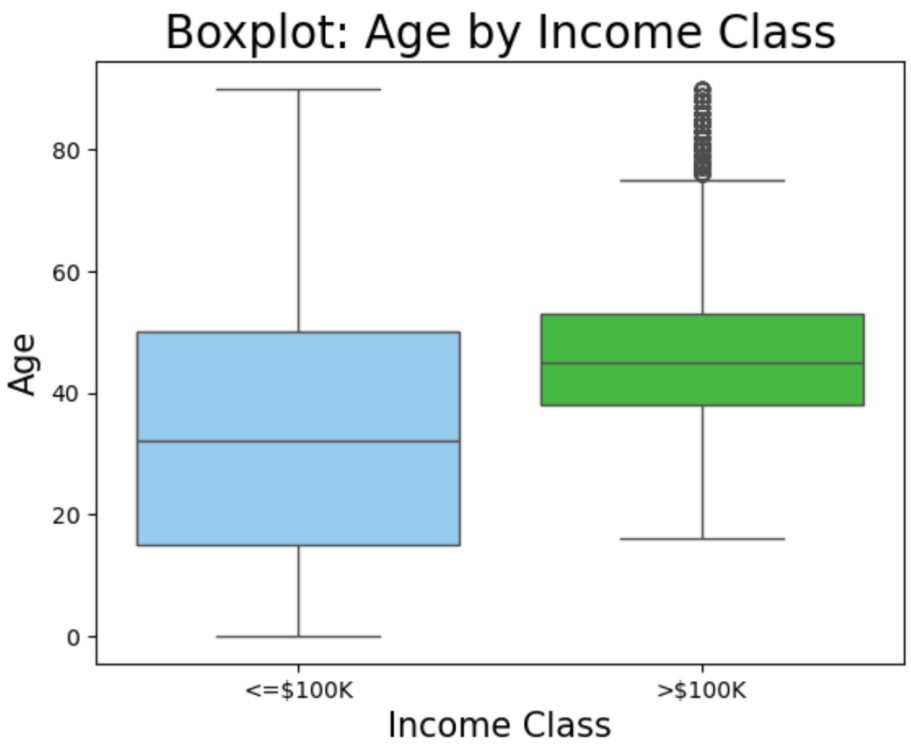


Point Biserial Correlation with Income: **0.26**

Weeks worked in year was converted into three categories:
not working, part-time, full-time.

Data Deep Dive: Distributions of Continuous Variables, Age



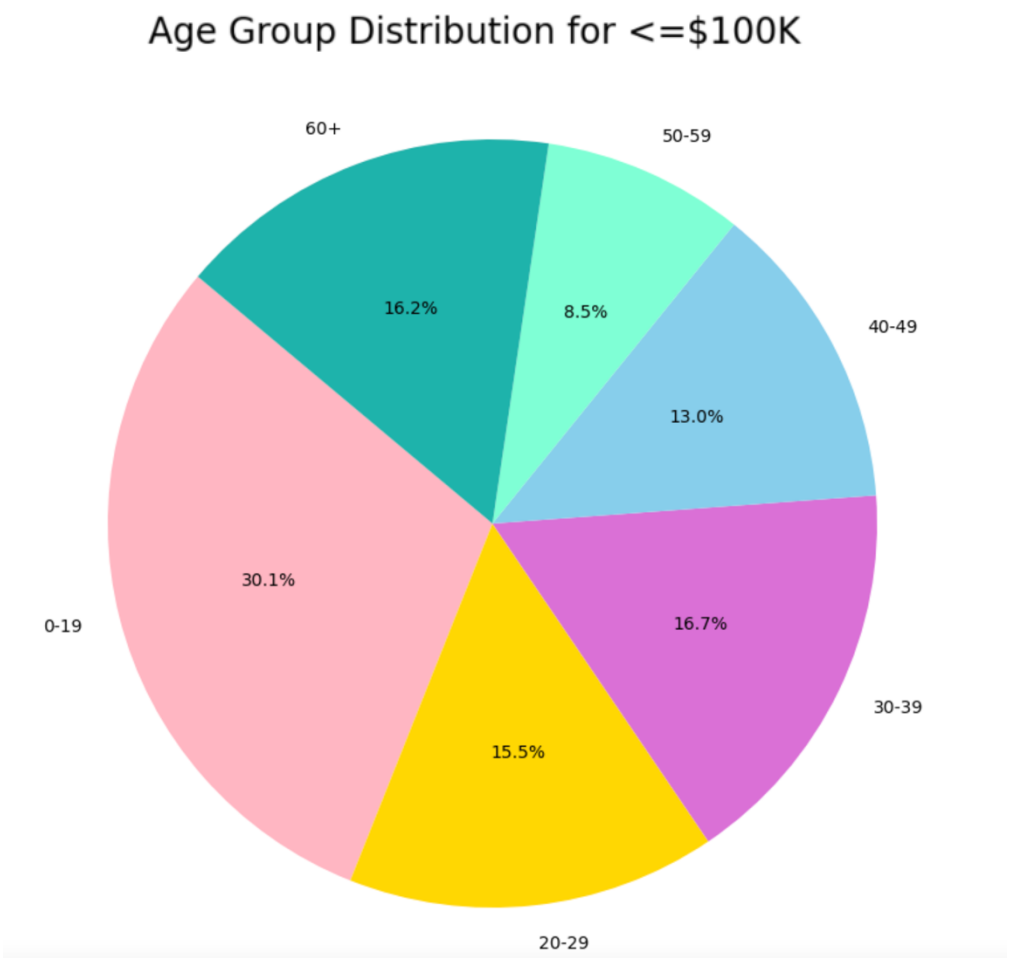


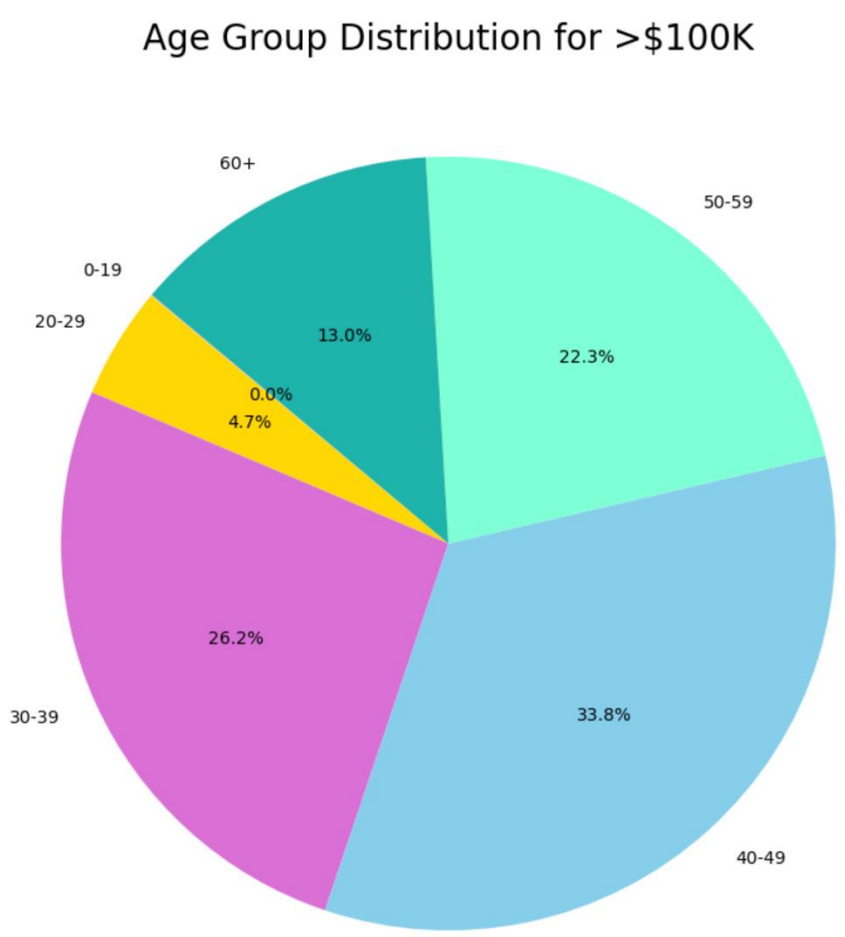
Point Biserial Correlation with Income: **0.13**

target	<=$100K							>$100K						
Quantile	**0.00**	**0.10**	**0.25**	**0.50**	**0.75**	**0.90**	**1.00**	**0.00**	**0.10**	**0.25**	**0.50**	**0.75**	**0.90**	**1.00**
age	0.0	6.0	15.0	32.0	50.0	68.0	90.0	16.0	32.0	38.0	45.0	53.0	62.0	90.0

Data Deep Dive: Distributions of Continuous Variables, Age *

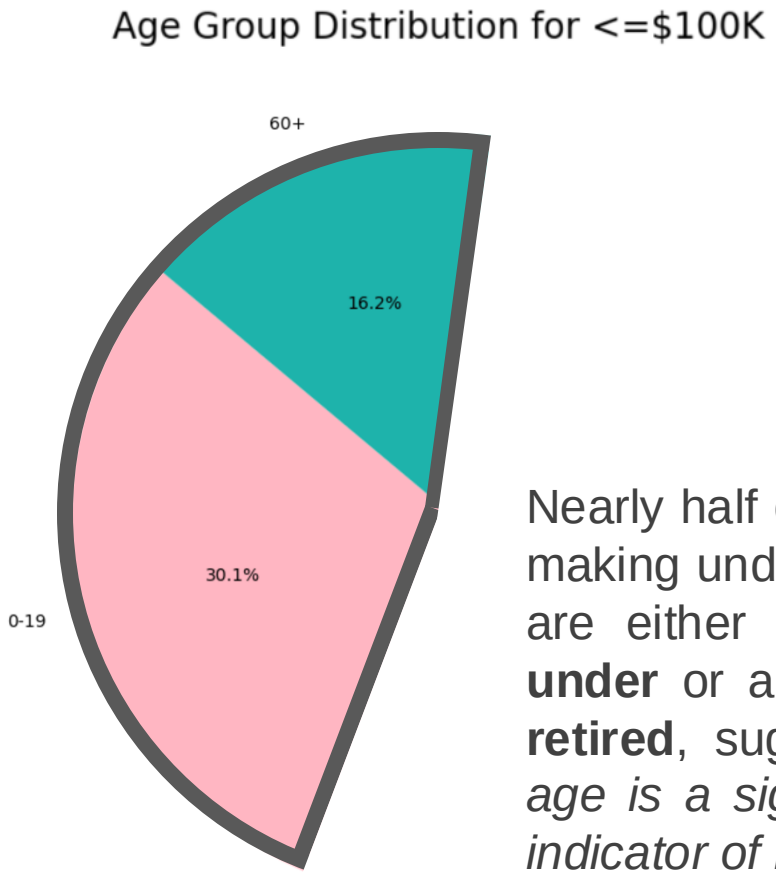
**Categorized here for illustrative purposes only.*



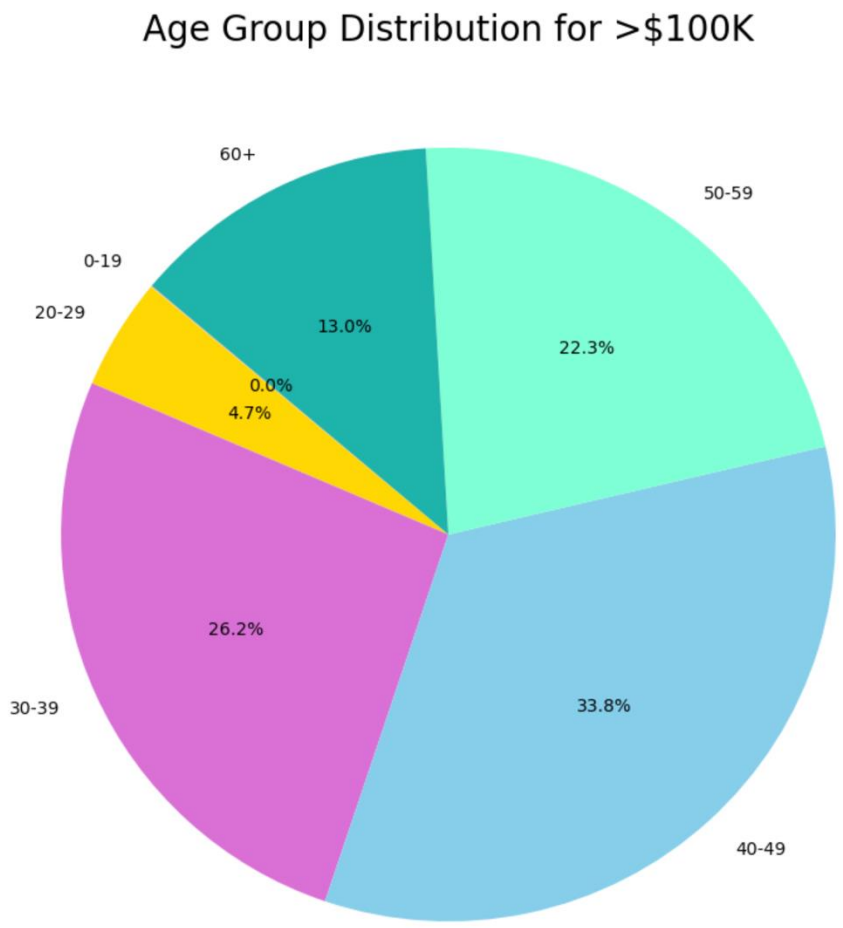


Data Deep Dive: Distributions of Continuous Variables, Age *

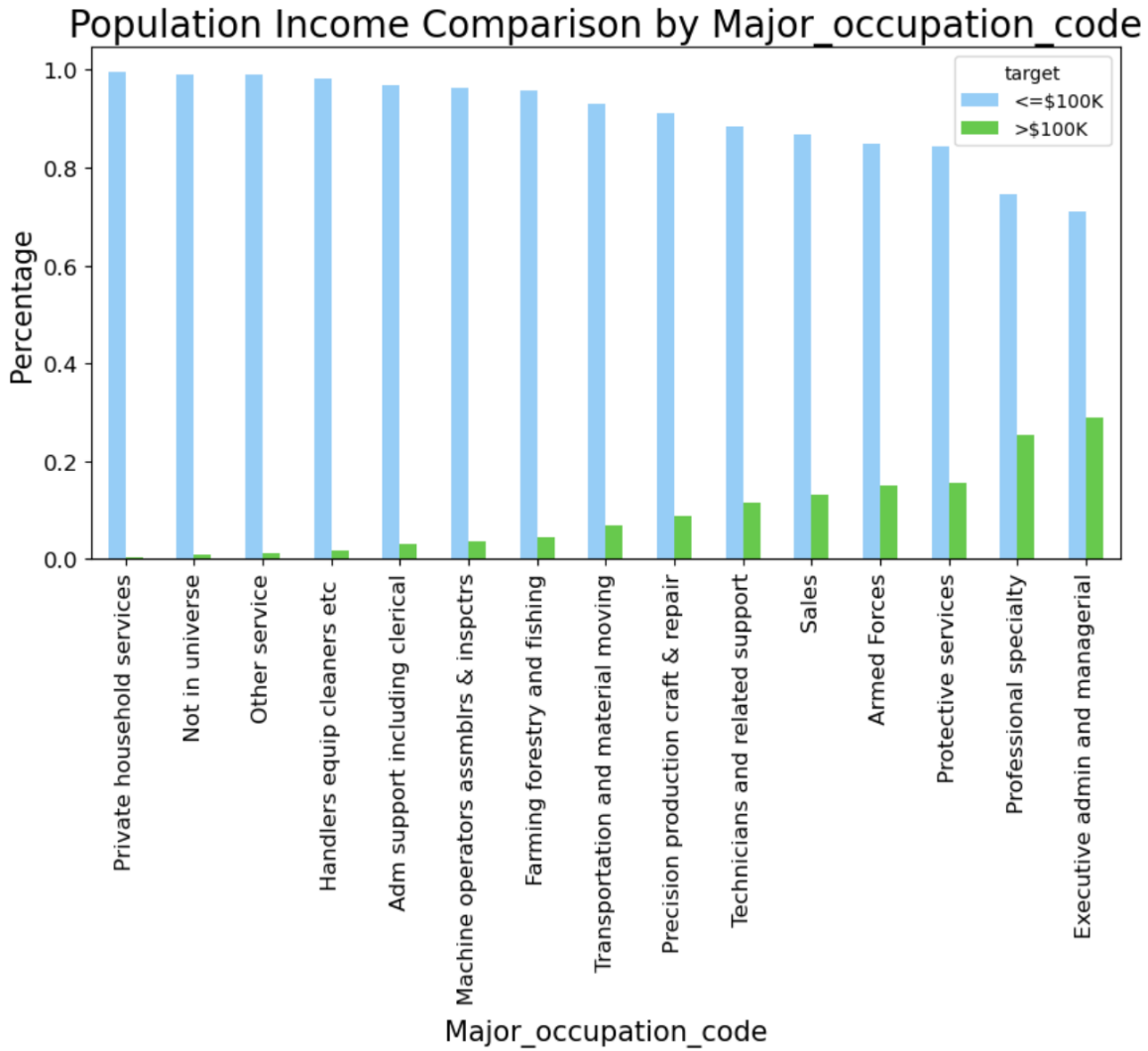
**Categorized here for illustrative purposes only.*



Nearly half of those making under $50K are either **19 and under** or are likely **retired**, suggesting *age is a significant indicator of income*.

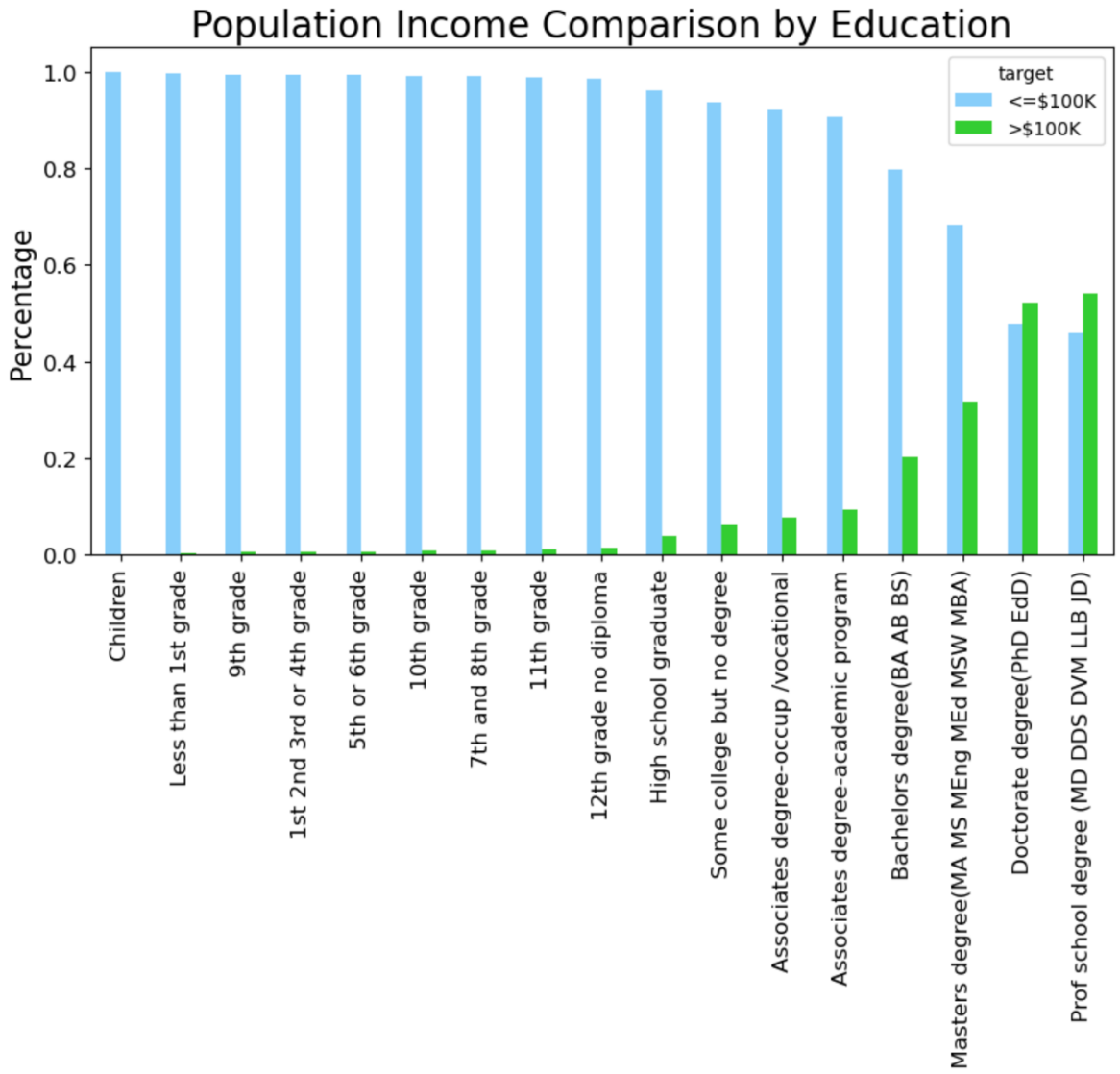


Data Deep Dive: Distributions of Categorical Variables, Occupation



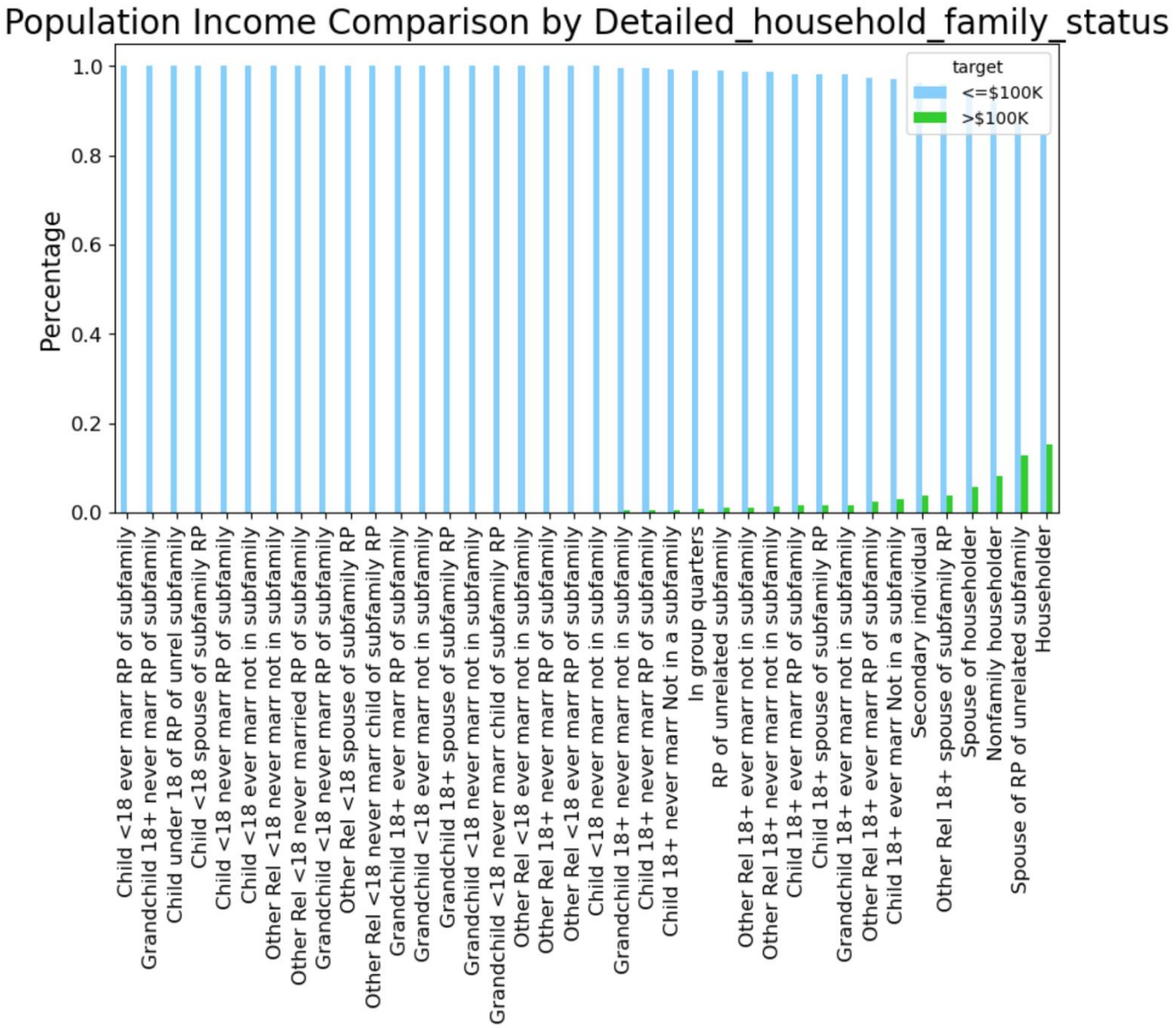
- Distribution exhibits a **clear interaction of *major occupation* with *income bracket*.**
- **Executive managerial work and professional specialties** are the most likely to make above $100K/year.

Data Deep Dive: Distributions of Categorical Variables, Education



- Distribution exhibits a **clear interaction of *education* with *income bracket*.**
- **Over half of all professional school degrees (e.g. MD) and doctorate degrees** make over $100K/year.

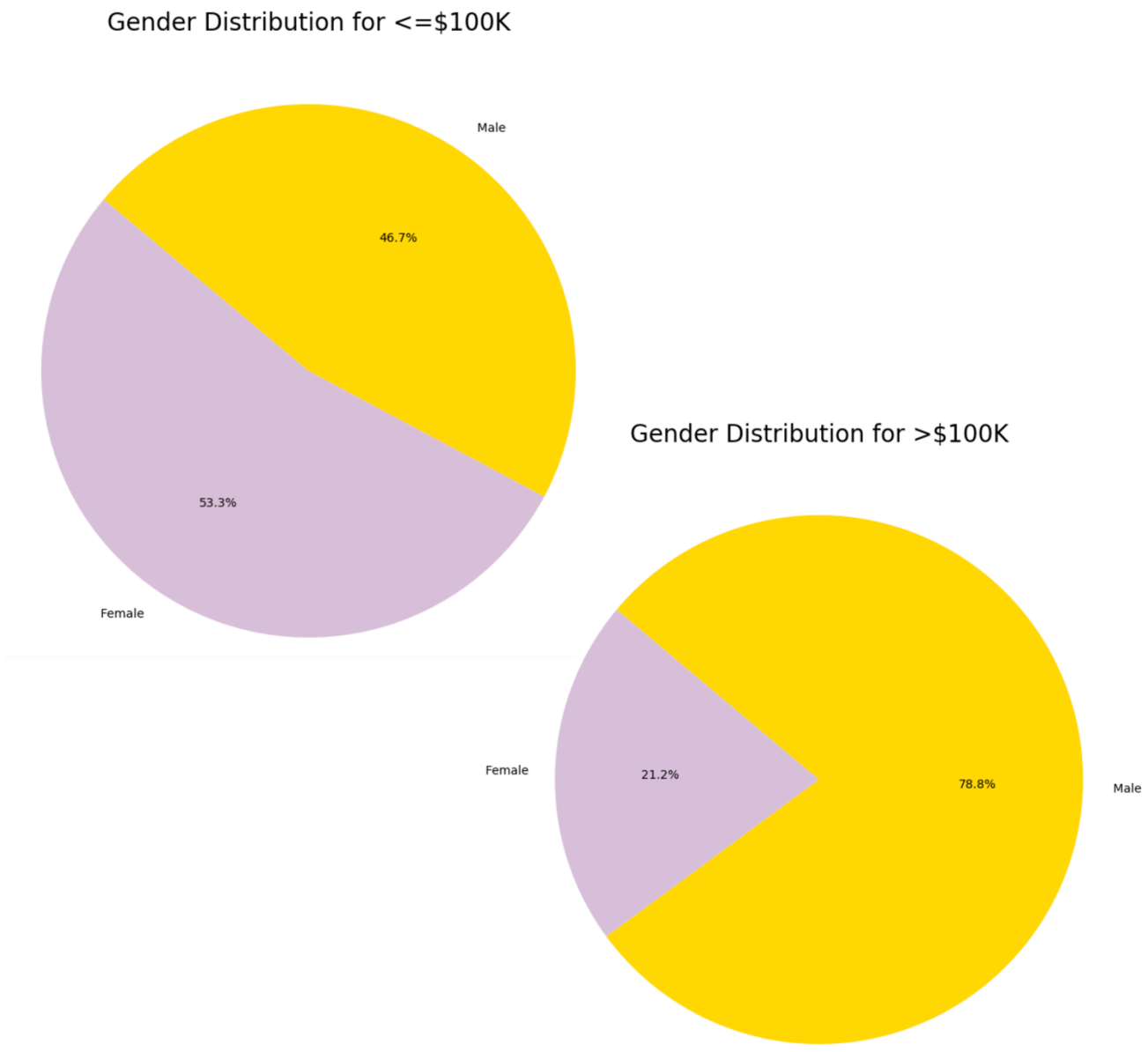
Data Deep Dive: Distributions of Categorical Variables, Household Family Status



Distribution exhibits a **clear interaction of *household family status* with *income bracket*.**

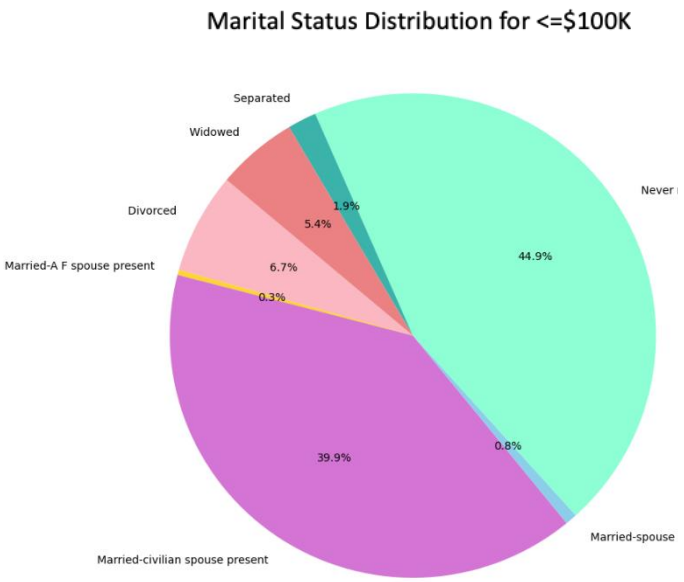
Heads of households are the most likely to earn above $100K.

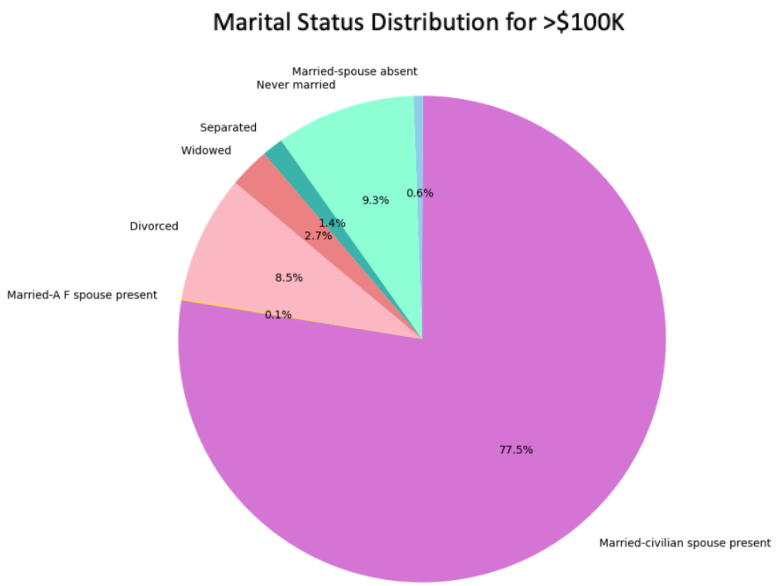
Data Deep Dive: Distributions of Categorical Variables, Gender

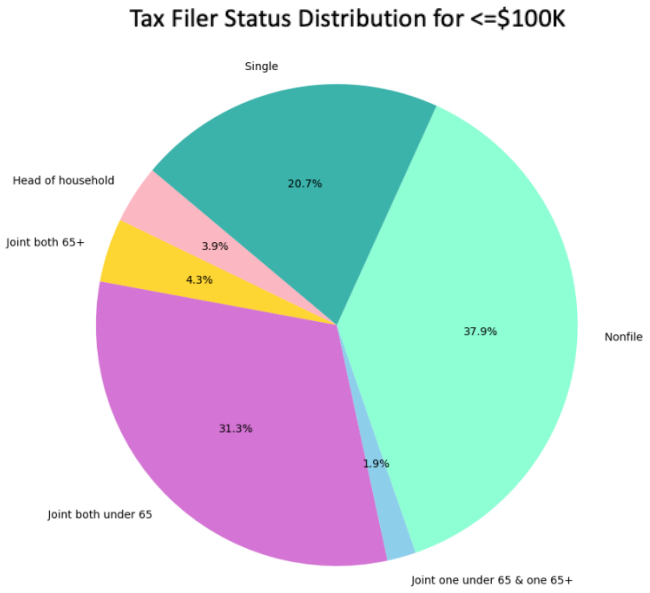


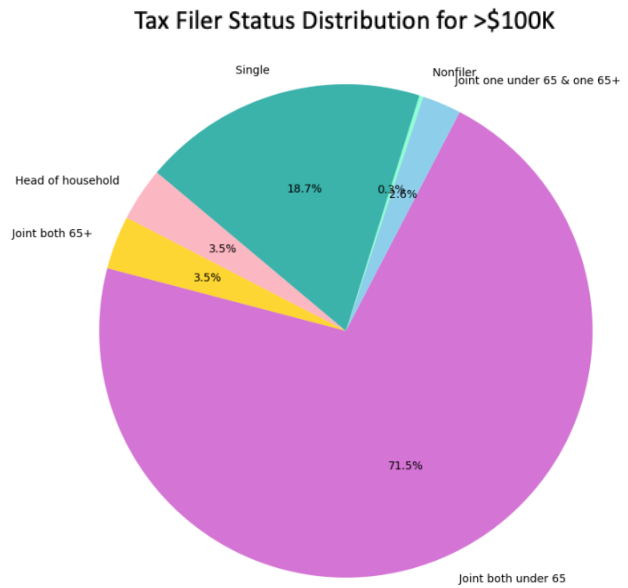
- Distribution exhibits a **clear interaction of *gender* with *income bracket*.**
- Men are nearly **4X as likely** to make over $100K.
- The **majority of the population** who makes under 100K are women.

Data Deep Dive: Distributions of Categorical Variables, Marital and Tax Status



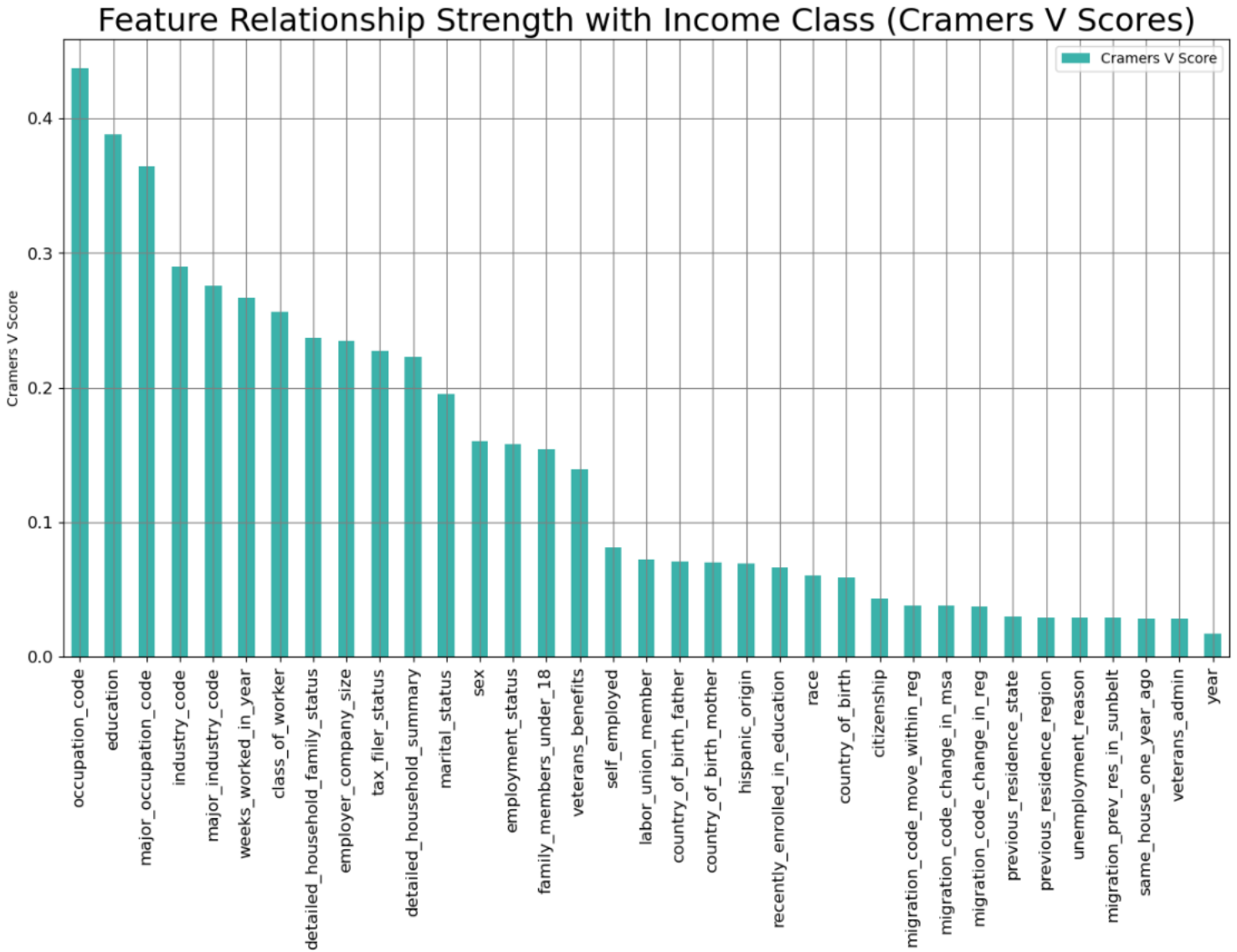






- Distributions exhibit **clear interactions of *marital status* with *income bracket*** and ***tax status* with *income bracket*.**
- Nearly **80%** of all people who make over $100K are **married**.
- Nearly **half** of people who make under $100K have **never been married**.

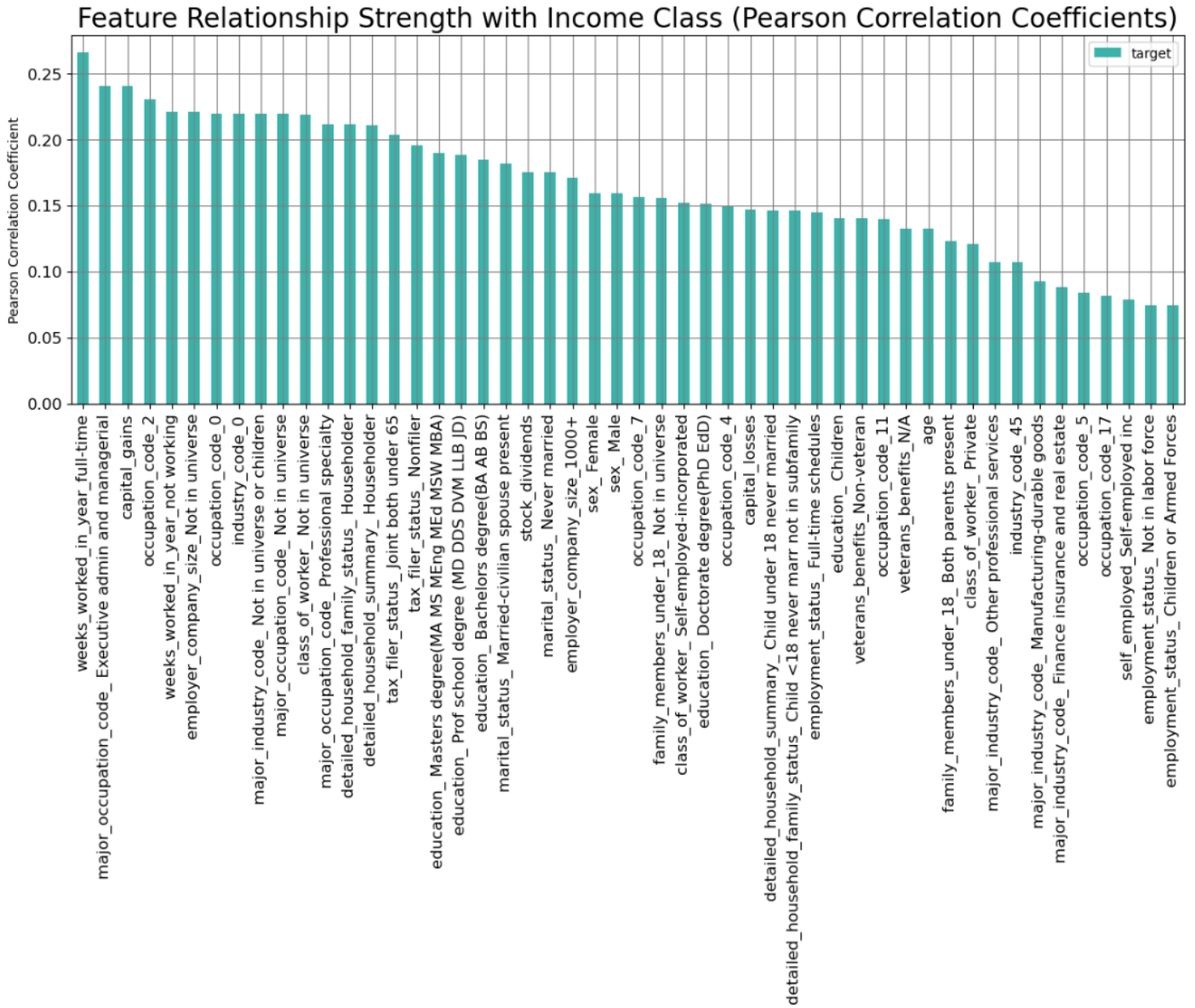
Data Deep Dive: Feature Relationship Strength with Income Class, Cramer's V Scores



- ***Occupation*** has a moderate-to-strong relationship with income.
- ***Education, industry, weeks worked in a year, class of worker, household family status and summary, employer company size,*** and ***tax filer status*** have a moderate relationship with income.
- ***Marital status, gender, employment status, family members under 18,*** and ***veterans benefits*** have weak relationships.

Data Deep Dive: Feature Relationship Strength with Income Class, Correlations

**Post-encoding into dummy variables.*



Pearson Correlation Coefficient >=0.15:

- ***Full-time work***
- ***Executive managerial occupation or professional specialty***
- ***Capital gains***
- ***Working v. not working***
- ***Being the head of household***
- ***Being married***
- ***Education, Bachelor's degree and beyond***
- ***Working for a very large company***
- ***Stock dividends***
- ***Gender***
- ***Kids at home***
- ***Being self-employed (incorporated)***

Model Construction & Assessment

Data Modeling: Machine Learning Prep

Feature Engineering

- ***Capital gains***, ***capital losses***, and ***stock dividends*** were converted into binary features (none or positive).
- ***Wage p. hour*** was converted into a binary variable of **greater / less than $50/hour.**
- ***Weeks worked in year*** was converted into three categories: **not working, part-time, full-time.**
- All other demographic variables were used as-is (post-cleaning) as features.
- Features with miniscule-to-none feature importance were **removed from machine learning models when their inclusion was counterproductive** (adding noise).

Training & Test Prep

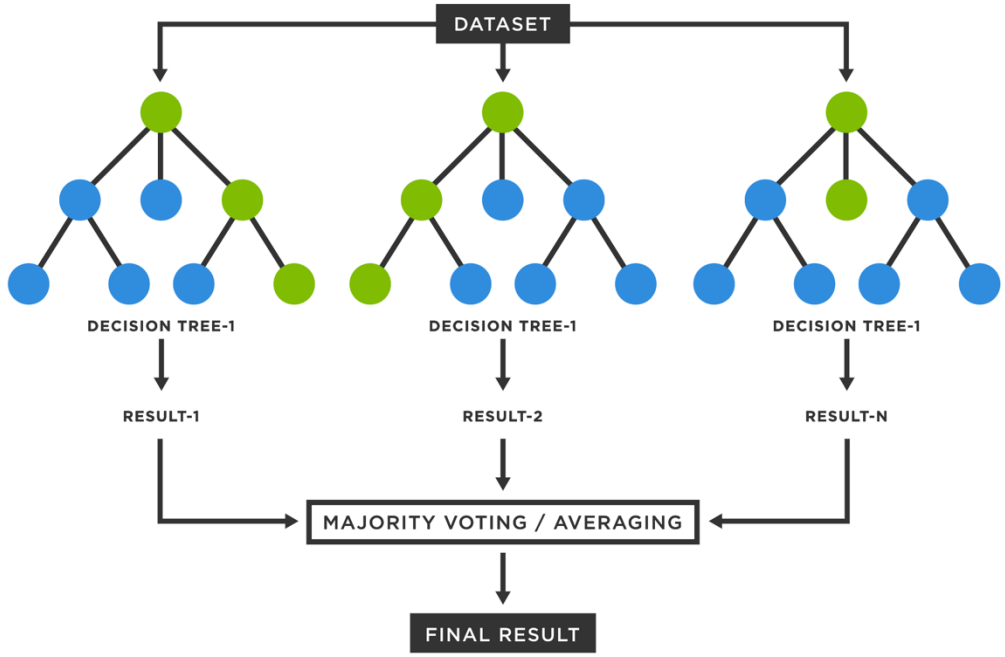
- Categorical features were encoded into **dummy variables** with OneHotEncoder, making for **517** features total.
- The **target variable was separated** from the features.
- Features were **scaled** using StandardScaler.
- Train: 183,912 rows
  Test: 96,256 rows
- Data was **resampled** to account for the class imbalance
    - Both **oversampling** and **synthetic** sampling were attempted.

Data Modeling: Random Forest, A Machine Learning Algorithm

Random Forest:

Random forest is a type of **decision tree model**. It works by splitting the data into subsets based on feature values, creating a tree-like structure where each **internal node** represents a decision on a **feature**, each **branch** represents an outcome of that decision, and each **leaf node** represents a final prediction or output.

It is an **ensemble method of supervised machine learning**. Ensemble methods combine predictions obtained from *multiple* base estimators to improve the overall prediction/robustness.



Data Modeling: Machine Learning Model Construction & Result

**Time constraints prevented thorough hyperparameter tuning.*

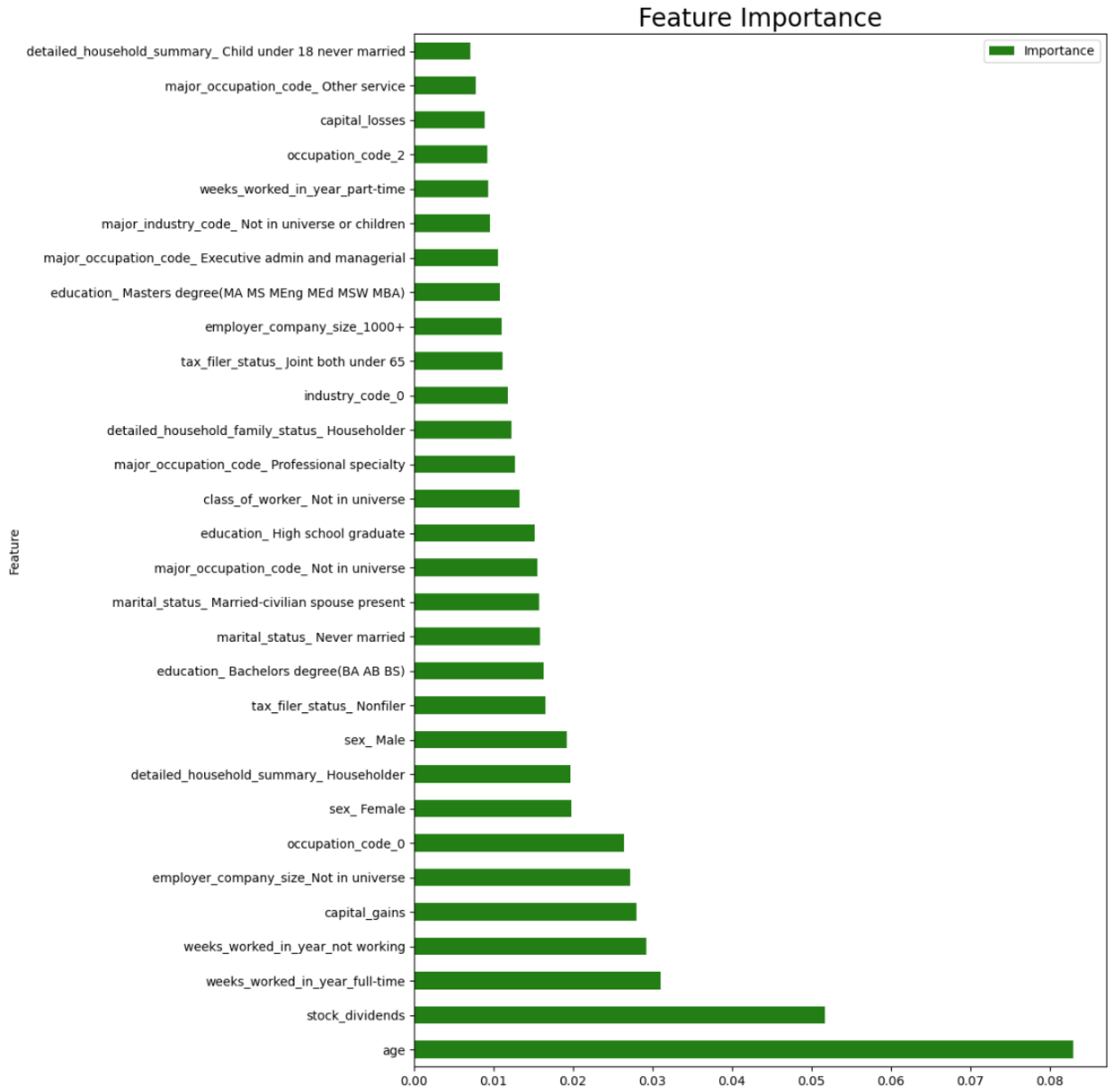
Random Forest Parameters & Result:

- ***100 estimators***
- ***Bootstrapped***
- ***Balanced*** *class weights*
- ***Strata instance weights were accommodated*** *via the 'sample weight' parameter*

	precision	recall	f1-score	support
<=$100K	0.96	0.99	0.97	88476
>$100k	0.71	0.36	0.48	6185
accuracy			0.95	94661
macro avg	0.83	0.68	0.73	94661
weighted avg	0.94	0.95	0.94	94661

ROC AUC Score: **0.93**

Random Forest Feature Importance



- ***Age*** is the most important feature followed by ***stock dividends***.
- ***Working full-time v. not working, capital gains,*** and ***occupation followed.***
- ***Gender, being the head of a household, not being a tax filer, having a Bachelor's degree v. a high school education,*** and ***marital status*** *all had similar feature importance.*
- ***Joint tax filer status, working for a very large employer, having a Master's degree, working in executive managerial roles, weeks worked per year, capital losses, and having children under 18 at home*** likewise were important.

Key Findings & Business Application

Summary of Key Findings from Random Forest Feature Importance

Machine learning model **feature importance scores** corroborate closely with **Point Biserial correlations**, **Cramer's V scores**, **Pearson correlations**, significant **Chi-squared** associations, and **feature/target interactions observed via plots**.

SUMMARY: *What predicts income?*

- ***Age.*** Being in a working age bracket is associated with higher income.
- ***Reporting stock dividends, capital gains, capital losses***
- ***Occupation matters***. Income is likely dependent on occupation, especially when you are in an executive or managerial role.
- ***Working full-time v. not working.*** Working generates earnings.
- ***Education.*** Bachelors and advanced degrees are associated with more income.
- ***Gender***. Men are more likely to earn more.
- ***Being the head of household***.
- ***Working for a very large company.***
- ***Being married and filing joint taxes.***
- ***Having children at home.***

Insights for Deep Sea Asset Management

BUSINESS APPLICATIONS

1. **Targeted Marketing for Financial Products**:
    A. **Age and Working Status**: Knowing that people in the working age bracket with full-time jobs have higher income potential, the company can focus its marketing efforts on working professionals between the ages of 30 and 55. This demographic is more likely to have disposable income and be interested in financial services.
    B. **Education and Occupation**: Since higher education and executive roles are associated with higher income, the company could target individuals with advanced degrees or specific occupations (e.g., executives, managers) for its high-end advisory services. This can be done by targeting ads on professional networks like LinkedIn, especially among users with these qualifications.

2. **Product Development and Pricing Strategy**:
    A. **Capital Gains and Stock Dividends**: The insight that reporting capital gains and stock dividends is associated with higher income could lead the company to develop specialized investment products or portfolios aimed at high-net-worth individuals. These products could focus on tax-efficient investments, targeting wealth clients who report stock dividends or capital gains to help them manage and grow their wealth more effectively.
    B. **Head of Household**: The company could create packages or advisory services tailored specifically for heads of households, who may be making more financial decisions for their families. These packages might include estate and legacy planning, education savings plans, and life insurance to secure their family's future.

3. **Client Retention Strategies**:
    A. **Gender**: Understanding that men might generally earn more could shape how the company approaches retention strategies. They might create specialized programs or incentives for male clients in their 30s and 40s who have been identified as high earners. Simultaneously, they could createtargeted outreach programs to support female clients in reaching their financial goals, as an opportunity to promote gender-specific financial planning services.
    B. **Having Children at Home**: The firm could design retention strategies focusing on clients with children at home by offering college savings plans, educational workshops on family financial planning, or loyalty programs that reward long-term clients who invest in family-oriented financial products.

4. **Choosing Where to Expand Services**:
    A. **Very Large Companies**: Since income is associated with working for large companies, the wealth management firm might consider partnerships with big corporations to attract clients who have more disposable income to invest. They could offer on-site or virtual seminars on financial wellness, investment options, and retirement planning for employees of these companies.

5. **Tax-Optimized Financial Advice**:
    A. **Married and Filing Joint Taxes**: The finding about marriage and filing status could prompt the company to offer more specialized tax-optimized financial advice for married couples. They might develop tax-efficient investment portfolios or consultative services to help couples maximize their joint income, save on taxes, and achieve financial goals together.

Questions? Happy to connect over email,
elle.nicole.roberts@gmail.com