ADULT INCOME CENSUS

Adam Benko, Peter Chen, Seunggyun Shin,
Courteney Chan, Jennie Hsu

01

Introduction

02

Data Understanding

03

Research Questions

04

Data Pre-processing

04

Modeling & Analysis

05

Conclusion

01 PROJECT INTRODUCTION

Data

- Kaggle Adult Income Census Data *(31,947 rows * 12 columns)*

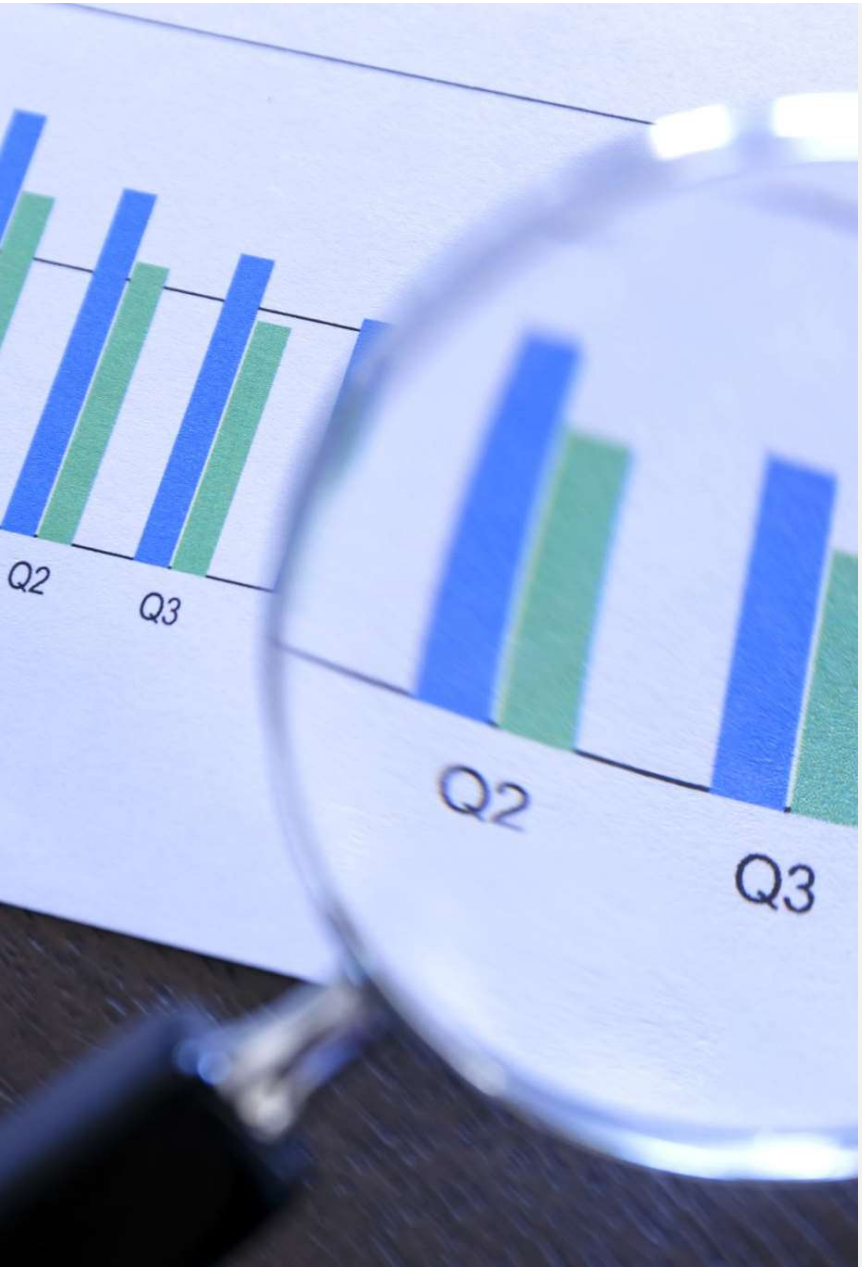
- https://www.kaggle.com/datasets/anaghakp/adult-income-census?resource=download

Background

- The dataset includes demographic features, and for each observation, it indicates whether the income exceeds 50K or not
- Important to observe social tendencies and identify inequalities
- Utilized in economic forecasting and workforce planning

Objective:

- Deliver insights corresponding with the research questions through visualizations and determine features that influence income level



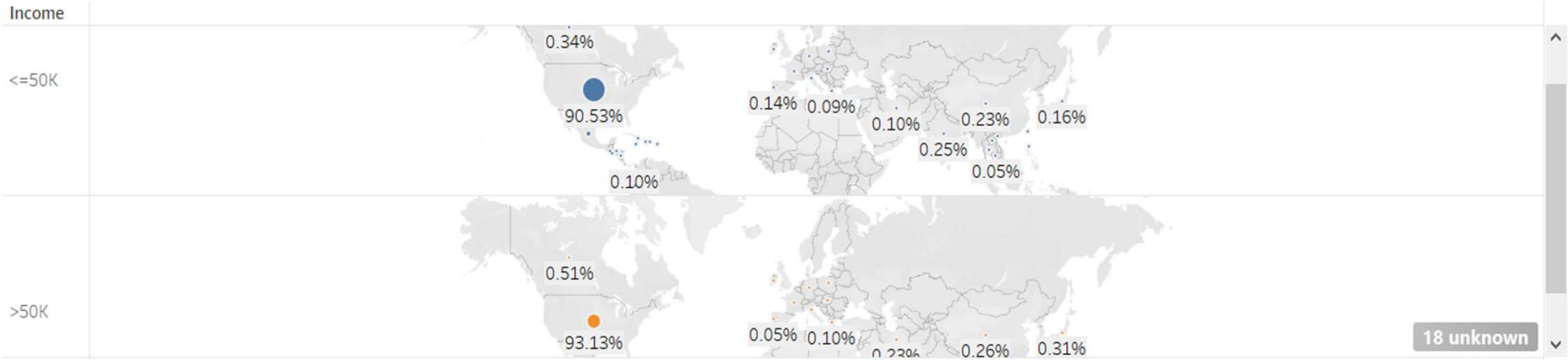
02 DATA UNDERSTANDING

- Observe data through Tableau visualizations to extract interesting research questions

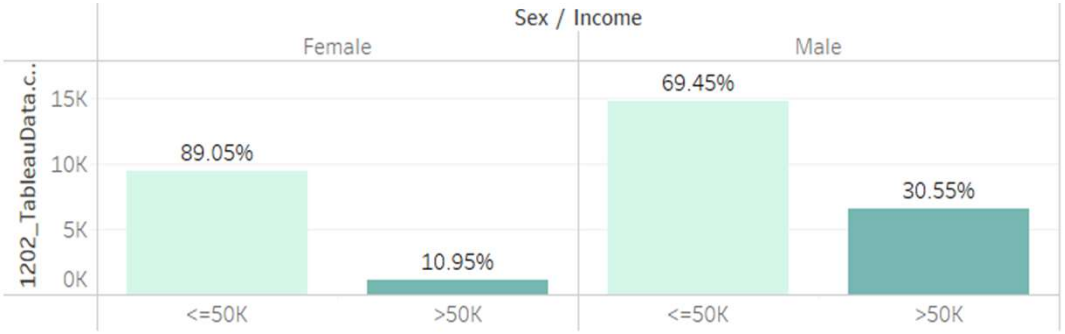
Statistic	Age	Education.num
Count	31,947	31,947
Mean	38.57	10.07
Std	13.65	2.56
Min	17.00	1.00
25%	28.00	9.00
50% (Median)	37.00	10.00
75%	48.00	12.00
Max	90.00	16.00

STORY POINT 1

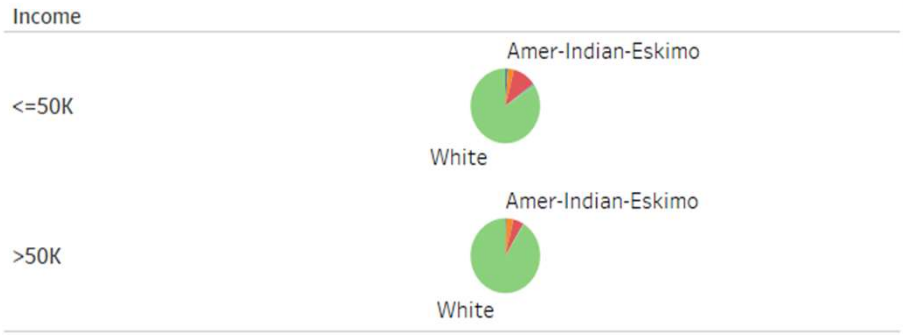
Native country: North America people account the largest proportion of the income.



Gender: Proportion of who make more than 50K per year by genders

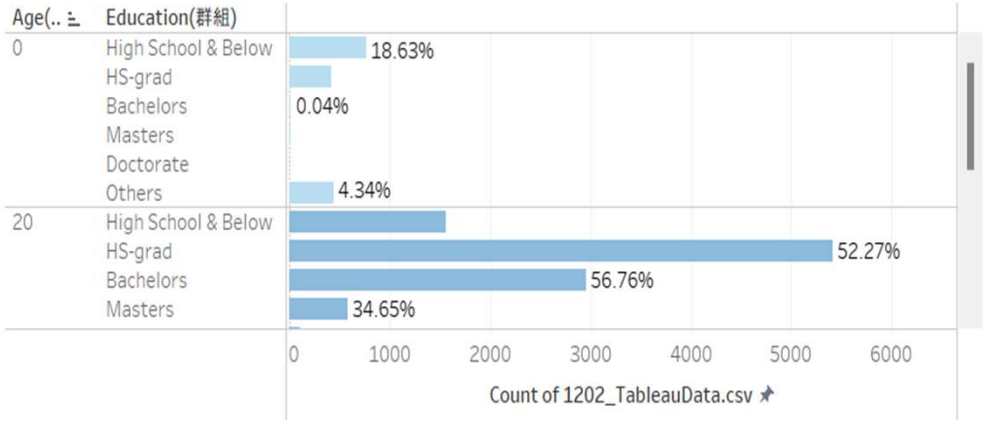


Race: White people make up the largest proportion of high income earners.

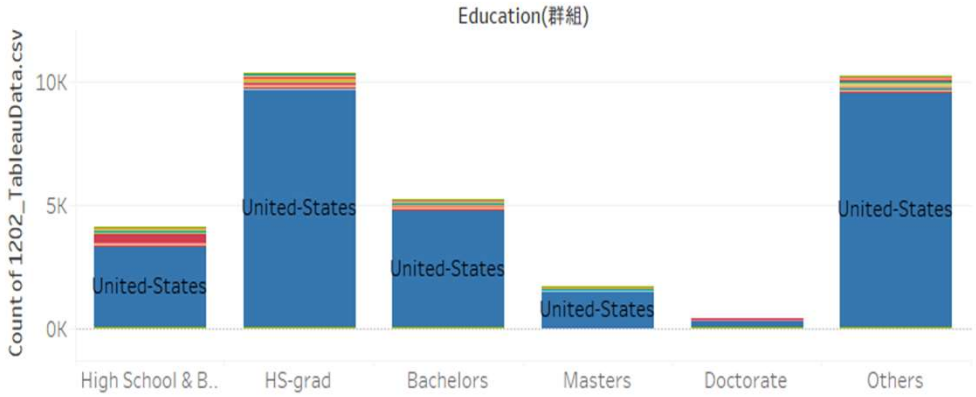


STORY POINT 2

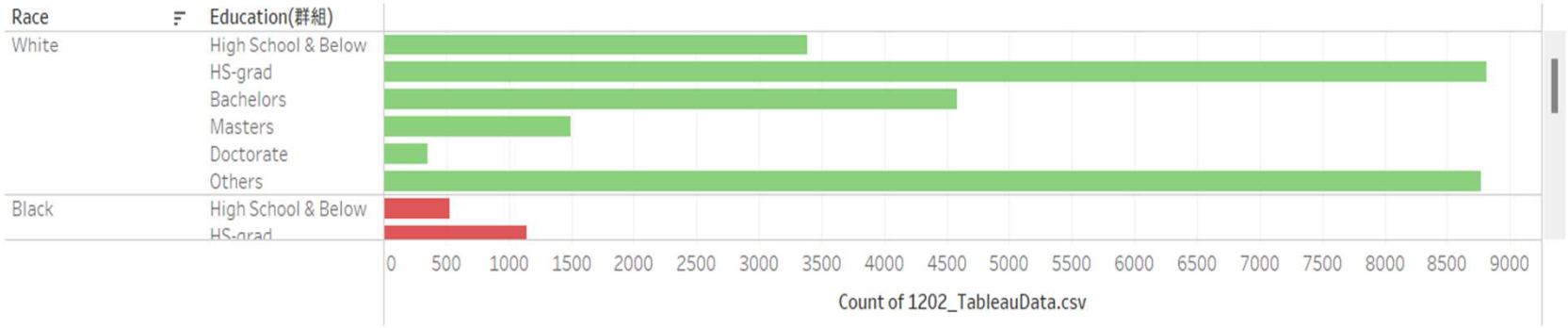
Age: People around 20-40 years tend to have the largest proportion of higher education level.



Native country: American residents tend to have the largest proportion of higher education level.



Race: White people significantly have higher levels of education attainment.



03 RESEARCH QUESTIONS

Key Questions

Research Q1

- Is there significant differences in income between gender, and is difference in education level causing the gap of income?

Research Q2

- Which factors contribute the most to income/ help us understand where income inequality stems from?

04 DATA PRE-PROCESSING

- Prepare Data for Modeling

1. Drop NA values (6% data loss)

```
In [6]: df = df.dropna(axis = 0)
        df.shape

Out[6]: (30162, 12)
```

2. Adjust column values: education, marital.status
   - Reduce dummy variables by observing group means

```
#education
def replace_education_level(education):
    # First condition for 'no_highschool_deg'
    if education in ['11th', '10th', '7th-8th', '9th', '12th', '5th-6th', '1st-4th', 'Preschool']:
        return 'no_highschool_deg'
    # Second condition for 'others'
    elif education in ['Assoc-voc', 'Assoc-acdm']:
        return 'Associate'
    elif education in ['Some-college']:
        return 'college degree'
    elif education in ['Doctorate', 'Prof-school']:
        return 'Doctor/Prof'
    # If none of the above conditions are met, return the original education value
    else:
        return education

df["education"] = df["education"].apply(replace_education_level)
```

3. Drop Columns & Create Dummies (Before modeling)

```
X = df2.drop(["income", "occupation", "relationship", "education", "native.country"], axis = 1)
X = pd.get_dummies(X, drop_first = True)
X = sm.add_constant(X)
Y = df2["income"].astype("float")
```

- Income: Response variable
- Relationship: Correlated w/ Marital Status
- Education: Correlated w/ Education Number
- Occupation: Correlated w/ Workclass
- Native Country: Biased towards the "United States"

05 MODELING &ANALYSIS

- Research Question 1

  - Is there significant differences in income between gender, and is difference in education level causing the gap of income?

1. Income by Gender

- One-tailed z-test for proportions ($\alpha = 0.05$)
- p: Proportion of who earn more than 50K

H0: $p_{male} \leq p_{female}$	H0: $p_{male} > p_{female}$
Z statistics	37.63
p-value	Almost 0

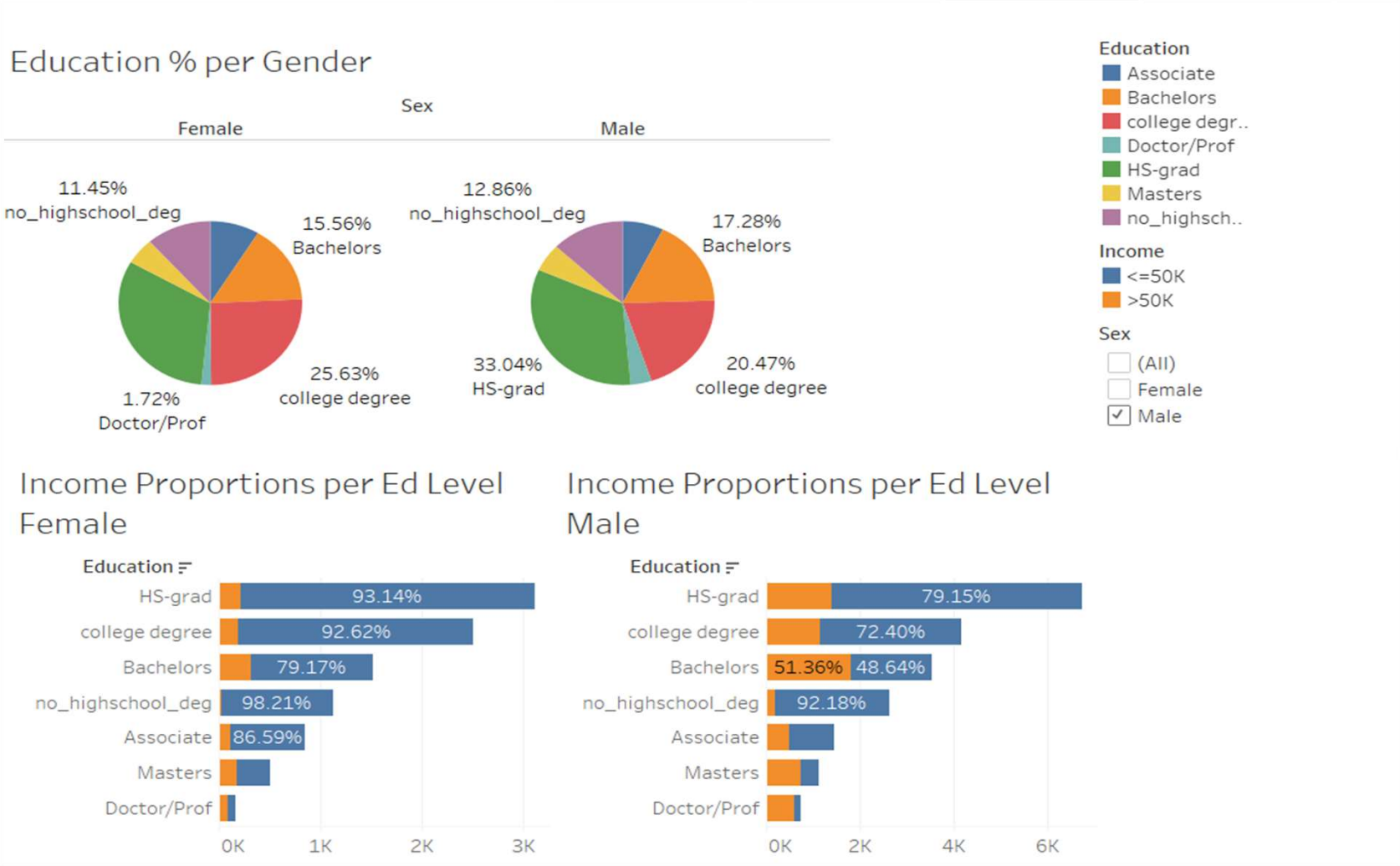
2. Education by Gender

- Two-tailed two sample t-test
- *Mu*: Mean of education numbers

H0: $Mu_{male} = Mu_{female}$	H0: $Mu_{male} \neq Mu_{female}$
t statistics	1.067
p-value	0.285

- There is a significant evidence that male earns more money than female does. However, as there is no evidence to claim that there is a difference in education level between two genders, education level is not a factor which causes the difference.

STORY POINT 3



05 MODELING &ANALYSIS

- Research Question II
  - Which factors contribute the most to income/ help us understand where income inequality stems from?
  - Selected models with high explainability

1. Logistic Regression
   - 82% fitted accuracy (Threshold: 0.5)
   - Age, Education, Marital Status

Dep. Variable:	income	**No. Observations:**	30162
Model:	Logit	**Df Residuals:**	30145
Method:	MLE	**Df Model:**	16
Date:	Mon, 04 Dec 2023	**Pseudo R-squ.:**	0.3201
Time:	02:11:17	**Log-Likelihood:**	-11507.
converged:	True	**LL-Null:**	-16925.
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-8.3922	0.249	-33.747	0.000	-8.880	-7.905
age	0.0274	0.001	18.580	0.000	0.024	0.030
education.num	0.4002	0.008	51.582	0.000	0.385	0.415
workclass_Local-gov	-0.6382	0.102	-6.251	0.000	-0.838	-0.438
workclass_Private	-0.5147	0.086	-5.982	0.000	-0.683	-0.346
workclass_Self-emp-inc	0.0924	0.113	0.821	0.412	-0.128	0.313
workclass_Self-emp-not-inc	-0.9277	0.099	-9.356	0.000	-1.122	-0.733
workclass_State-gov	-0.8332	0.116	-7.201	0.000	-1.060	-0.606
workclass_Without-pay	-369.5602	4.79e+79	-7.71e-78	1.000	-9.4e+79	9.4e+79
marital.status_married	2.5895	0.057	45.499	0.000	2.478	2.701
marital.status_married - absent	0.5138	0.208	2.474	0.013	0.107	0.921
marital.status_separated	0.5697	0.071	8.080	0.000	0.432	0.708
race_Asian-Pac-Islander	0.2570	0.227	1.133	0.257	-0.187	0.702
race_Black	0.3192	0.215	1.482	0.138	-0.103	0.742
race_Other	-0.3099	0.336	-0.921	0.357	-0.969	0.349
race_White	0.5521	0.206	2.679	0.007	0.148	0.956
sex_Male	0.3097	0.046	6.766	0.000	0.220	0.399

05 MODELING &ANALYSIS

Logistic Regression Model Findings:

- **Significant Predictors**: by p-values less than 0.05

1. Education: Important for income level, but not gender
2. Marital status (Married): Married females tend to earn more
3. Sex: Men are more likely to earn >50k more than women

05 RESEARCH FINDINGS

Logistic Regression Model Findings:

- **Model Statistics**:

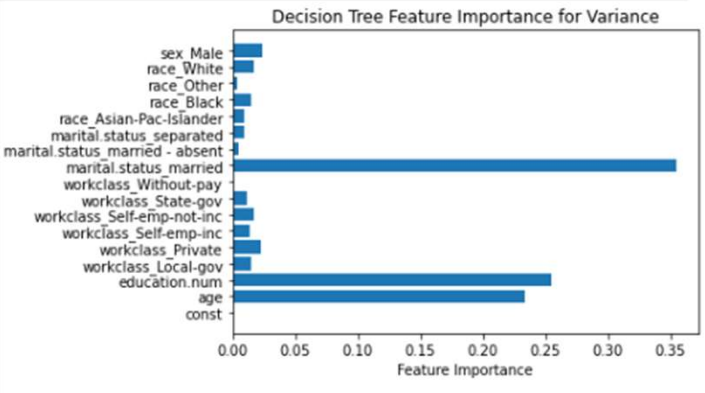
  1. pseudo-R-squared value: 0.3201: 32% of the variability in the income is explained by the model
  2. The LLR p-value: close to 0

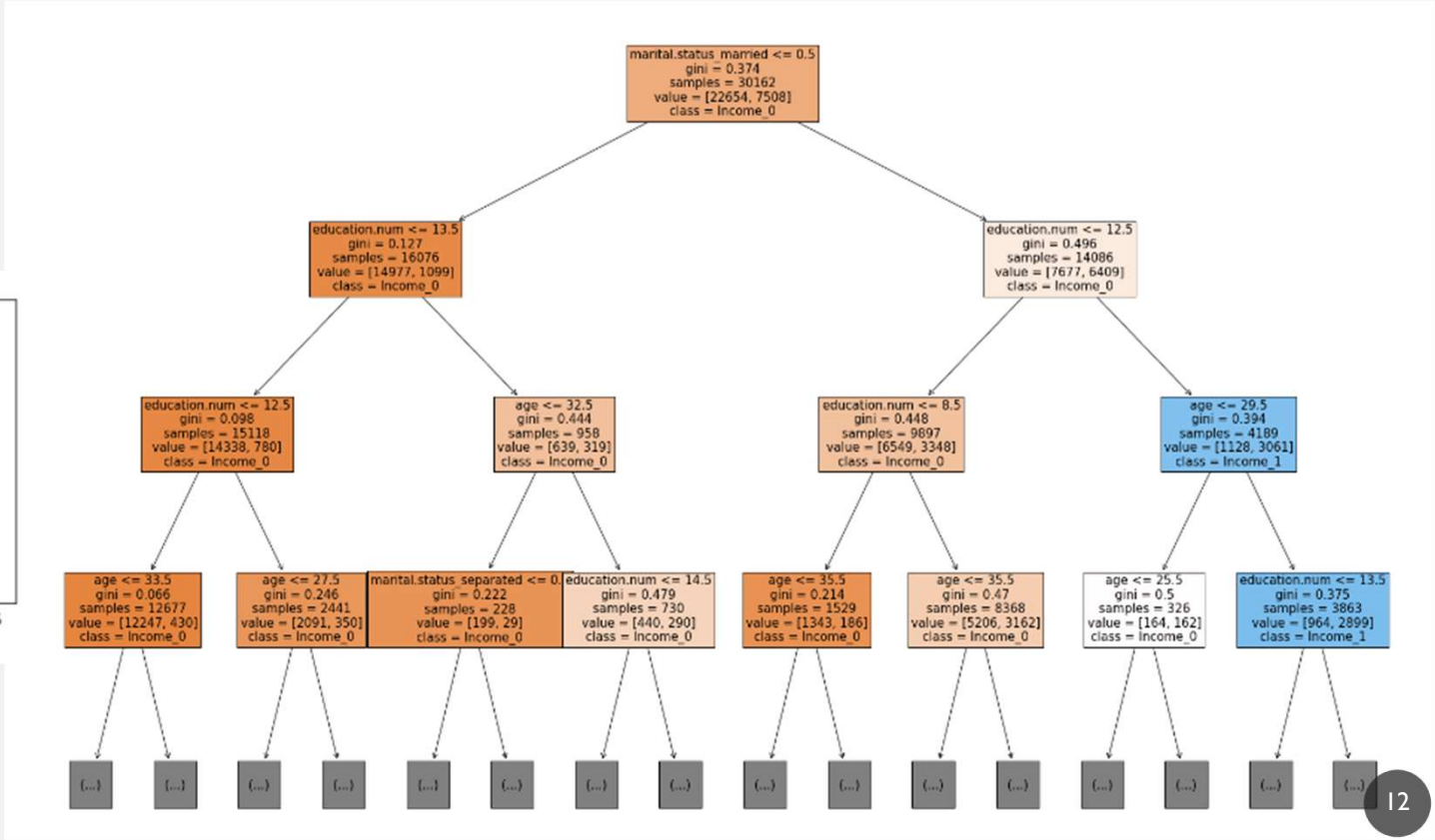
     statistically significant when compared to a null model with no predictors

05 MODELING &ANALYSIS

- Research Question II
  - Which factors contribute the most to income/ help us understand where income inequality stems from?

1. Decision Tree Classifier
   - 87% fitted accuracy
   - Age, Education, Marital Status





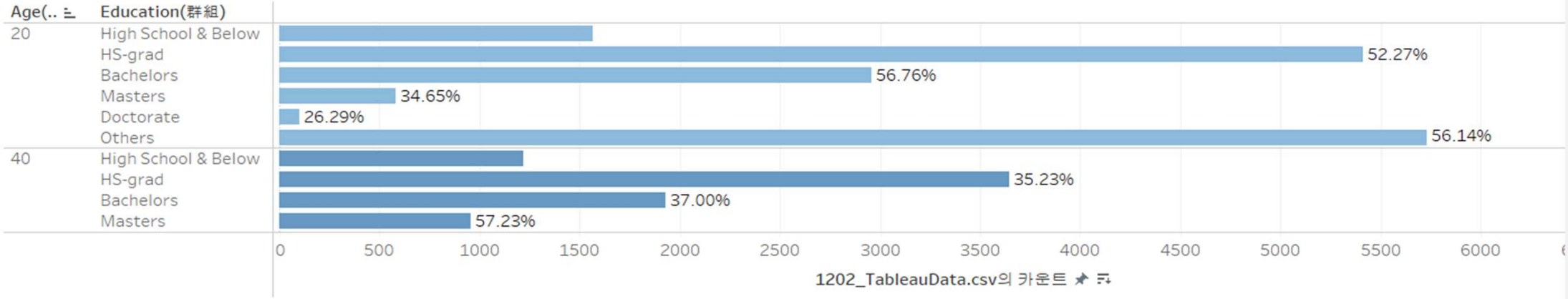
05 RESEARCH FINDINGS

- **Decision Tree Model Findings:**

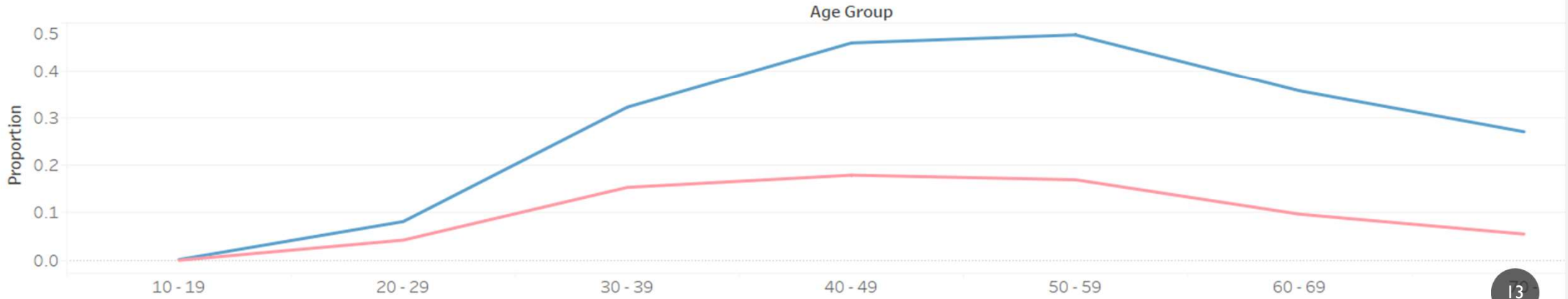
1. **Primary Splits**: the number of education years and marital status
2. **Thresholds for Splits**:
    - **Education**: <= 13.5 years and <= 12.5
    - **Age**: <= 29.5 years, <= 33.5 years, and <= 35.5 years, etc.
3. **Sample Sizes and Class Predictions**: Each node provides sample sizes and the class prediction (income 0 or income 1) for the data points that fall into that node, based on the splits made by the tree.

STORY POINT 4

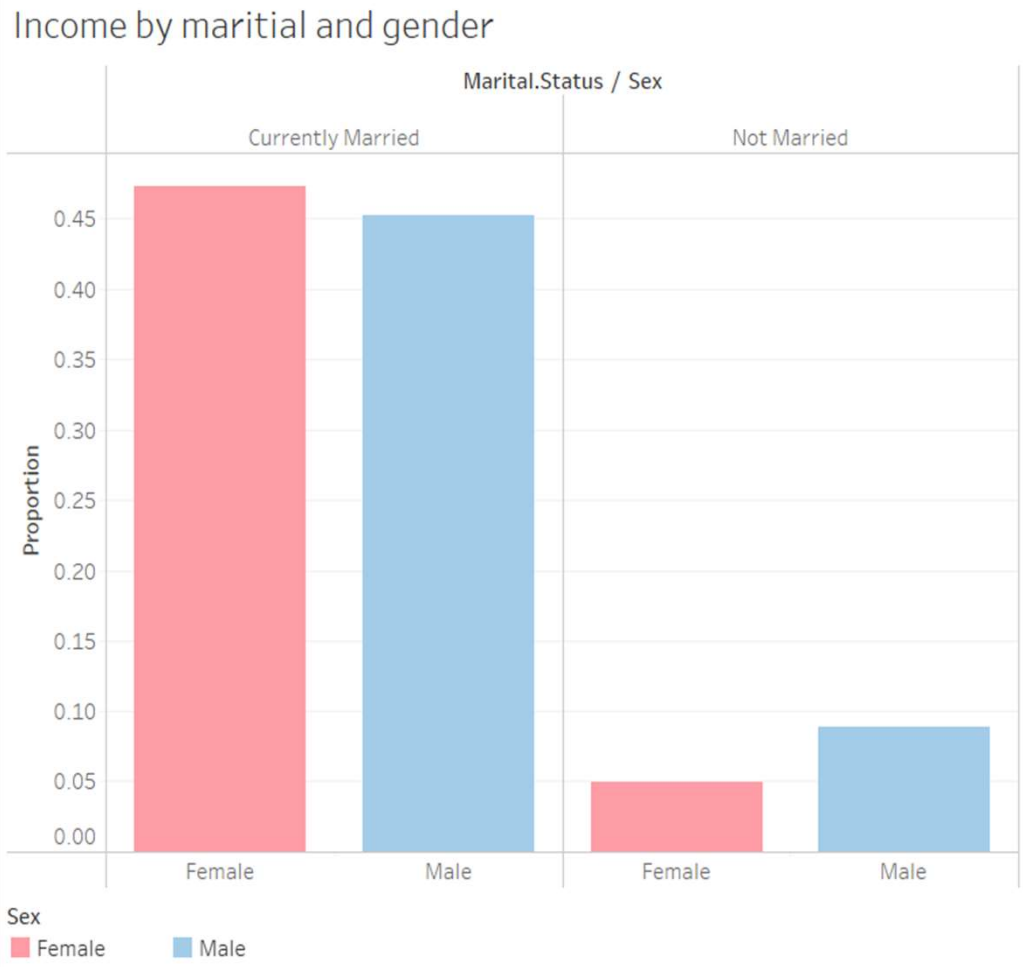
Age: People around 20-40 years tend to have the largest proportion of higher education level.

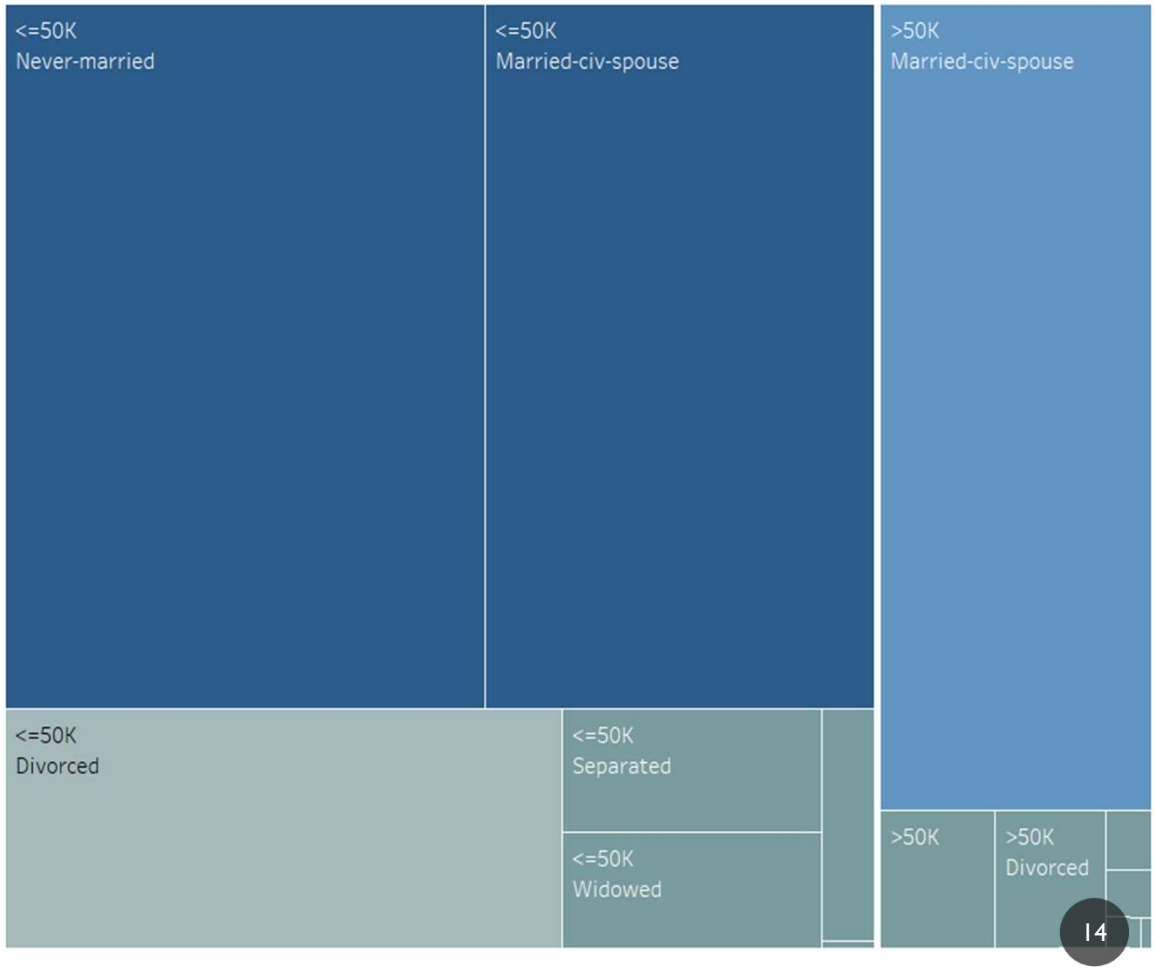


Proportion of people who earn more than 50K per year by Age groups



STORY POINT 5





06 PROJECT CONCLUSION

Summary

- There is a sufficient difference in income level by genders.
- Education level, age, and marital status are significant variables that has notable relationships with income level. However, they are not a factor causing income gap between genders.

Improvements

- The dataset is biased that there can be more reliable result with a better dataset.

  ex) Married proportion of female <<<< proportion of male

- The variable "fnlwgt", describes the weight of each row, can be utilized for more accurate analysis.