Detecting Affluence Rate on Census Dataset

Cesar Malenab

Regina Flores

Emmanuel Pedernal

Table of Contents

Data Preprocessing

Source

From UCI Machine learning repository

https://archive.ics.uci.edu/dataset/2/adult

Created by: Barry Becker and Ronny Kohavi

DOI: 10.24432/C5XW20

Contains census information from 1994 (USA)

The dataset is a fairly large set, consisting of 48,842 instances.

There are 14 attributes prescribed to each person:

income (‘>50K’ or ‘<=50K’), age, workclass
fnlwgt, education, education-num, marital-status
occupation, relationship, race, sex,capital-gain
capital-loss, hours-per-week, and native-country

Data Dictionary

Variable Name	Type	Demographic	Description	Missing val
Age	Integer	Age	N/A	No
Workclass	Categorical	Income	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.	Yes
Fnlwgt	Integer	N/A	N/A	No
Education	Categorical	Education Level	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.	No

Data Dictionary

Variable Name	Type	Demographic	Description	Missing val
education-num	Integer	Education Level	N/A	No
Marital Status	Categorical	Other	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.	No
Occupation	Categorical	Other	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners,	No
Relationship	Categorical	Other	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.	No

Data Dictionary

Variable Name	Type	Demographic	Description	Missing val
Race	Categorical	Race	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.	No
Sex	Binary	Sex	Female, Male	No
capital-gain	Integer	N/A	N/A	No
capital-loss	Integer	N/A	N/A	No
Native-country	Categorical	N/A	List of Countries	Yes
hours-per-week	Integer	N/A	N/A	No
Income	Binary	Income	>50K, <=50K.	No

Data Collection

The Team utilized UCI Machine Learning Repository to fetch the dataset directly

```
# Fetch dataset
from ucimlrepo import fetch_ucirepo, list_available_datasets
adult_raw = fetch_ucirepo(id=2)

X = adult_raw.data.features
y = adult_raw.data.targets

adult = pd.concat([adult_raw.data.features, adult_raw.data.targets], axis=1)
adult.head()
```

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	income
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K

Data Completeness

The Data Science team created a function that for each column in the dataset, it captures essential information such as unique values, cardinality (number of unique values), and the percentage of missing or unknown values. This facilitates informed decision-making by highlighting potential data quality issues and allowing for a nuanced understanding of the dataset

```
#get no of nunique, cardinality, % of NaNs and Uknowns
def analyse_cats(df, cat_cols):
    d = pd.DataFrame()
    cl = []; u=[]; s=[]; nans=[]; unknown=[]
    for c in cat_cols:
        cl.append(c);
        u.append(df[c].unique());
        s.append(df[c].unique().size);
        nans.append(round((df[c].isnull().sum()/(len(adult))*100), 2))
        unknown.append(round((df[c].astype(str).str.strip().eq("?").sum()/(len(adult))*100), 2))
    d["Feature"] = cl; d["Uniques"] = u; d["Cardinality"] = s; d["% Share of Nans"] = nans; d["% Share of Unknown (?)"] = unknown
    d["% Share of Nans or Unknowns (?)"] = d["% Share of Nans"] + d["% Share of Unknown (?)"]
    return d
```

	Feature	Uniques	Cardinality	% Share of Nans	% Share of Unknown (?)	% Share of Nans or Unknowns (?)
0	age	[39, 50, 38, 53, 28, 37, 49, 52, 31, 42, 30, 23, 32, 40, 34, 25, 43, 54, 35, 59, 56, 19, 20, 45, 22, 48, 21, 24, 57, 44, 41, 29, 18, 47, 46, 36, 79, 27, 67, 33, 76, 17, 55, 61, 70, 64, 71, 68, 66, 51, 58, 26, 60, 90, 75, 65, 77, 62, 63, 80, 72, 74, 69, 73, 81, 78, 88, 82, 83, 84, 85, 86, 87, 89]	74	0.00	0.00	0.00
1	workclass	[State-gov, Self-emp-not-inc, Private, Federal-gov, Local-gov, ?, Self-emp-inc, Without-pay, Never-worked, nan]	10	1.97	3.76	5.73
2	fnlwgt	[77516, 83311, 215646, 234721, 338409, 284582, 160187, 209642, 45781, 159449, 280464, 141297, 122272, 205019, 121772, 245487, 176756, 186824, 28887, 292175, 193524, 302146, 76845, 117037, 109015, 216851, 168294, 180211, 367260, 193366, 190709, 266015, 386940, 59951, 311512, 242406, 197200, 544091, 84154, 265477, 507875, 88506, 172987, 94638, 289980, 337895, 144361, 128354, 101603, 271466, 32275, 226956, 51835, 251585, 109832, 237993, 216666, 56352, 147372, 188146, 59496, 293936, 149640, 116632, 105598, 155537, 183175, 169846, 191681, 200681, 101509, 309974, 162298, 211678, 124744, 213921, 32214, 212759, 309634, 125927, 446839, 276515, 51618, 159937, 343591, 346253, 268234, 202051, 54334, 410867, 249977, 286730, 212563, 117747, 226296, 115585, 191277, 202683, 171095, 249409, ...]	28523	0.00	0.00	0.00

Data Binning

Selected features are binned from the dataset due to High cardinality attributes.

*age, workclass, education, occupation, hours-per-week, native-country

Age Bin:

- Under 18 years: Under 5 years, 5 to 17 years
- 18 to 44 years: 18 to 24 years, 25 to 44 years
- 45 to 64 years: 45 to 54 years, 55 to 64 years
- 65 years and over: 65 to 74 years, 75 to 84 years, 85 years and over

Work Class Bin:

- Private Sector Employee
- Government Employee
- Self-Employed Or Other

	count
age	48842.0
workclass	47879
fnlwgt	48842.0
education	48842
education-num	48842.0
marital-status	48842
occupation	47876
relationship	48842
race	48842
sex	48842
capital-gain	48842.0
capital-loss	48842.0
hours-per-week	48842.0
native-country	48568
income	48842

Data Binning

Selected features are binned from the dataset due to High cardinality attributes.

age, workclass, education, occupation, hours-per-week, native-country

Education Bin:

- No Schooling Completed
- Nursery or Pre-school Through Grade 12
- High School Graduate
- College or Some College
- After Bachelor's Degrees

Hours Bin:

- '0-20 hours'
- '20-40 hours'
- '40-60 hours'
- '>60 hours'

	count
age	48842.0
workclass	47879
fnlwgt	48842.0
education	48842
education-num	48842.0
marital-status	48842
occupation	47876
relationship	48842
race	48842
sex	48842
capital-gain	48842.0
capital-loss	48842.0
hours-per-week	48842.0
native-country	48568
income	48842

Data Binning

Selected features are binned from the dataset due to High cardinality attributes.

age, workclass, education, occupation, hours-per-week, native-country

Occupation Bin:

- Management, business, and financial
- Professional and related
- Service
- Sales and related
- Office and administrative support
- Farming, fishing, and forestry
- Construction and extraction
- Installation, maintenance, and repair
- Production
- Transportation and material moving
- Military specific

	count
age	48842.0
workclass	47879
fnlwgt	48842.0
education	48842
education-num	48842.0
marital-status	48842
occupation	47876
relationship	48842
race	48842
sex	48842
capital-gain	48842.0
capital-loss	48842.0
hours-per-week	48842.0
native-country	48568
income	48842

Data Reduction

Selected features are dropped from the dataset due to;

education-num : ordinal equivalent of education

fnlwgt : high cardinality, final weight assigned by survey

capital-gain, capital-loss : high cardinality, highly skewed, no context given

relationship: repetitive with marital-status and sex (ie. Husband = Married + Male)

```
[ ]  adult.drop(columns=["fnlwgt", "education-num", "capital-gain", "capital-loss", "native-country", "relationship"],
                inplace=True)
```

	count
age	48842.0
workclass	47879
fnlwgt	48842.0
education	48842
education-num	48842.0
marital-status	48842
occupation	47876
relationship	48842
race	48842
sex	48842
capital-gain	48842.0
capital-loss	48842.0
hours-per-week	48842.0
native-country	48568
income	48842

Final Attributes

```
# Describe all columns
adult.describe(include="all").T
```

	count	unique	top	freq
race	48842	5	White	41762
sex	48842	2	Male	32650
income	48842	4	<=50K	24720
age-group	48842	8	25 to 44 years	24770
workclass-group	48842	4	Private Sector Employee	33906
education-group	48842	4	College or Some College	22565
occupation-group	48842	11	Office and administrative support	7057
marital-status-group	48842	4	Married	22416
weekly-work-hours	48842	4	20-40 hours	30037

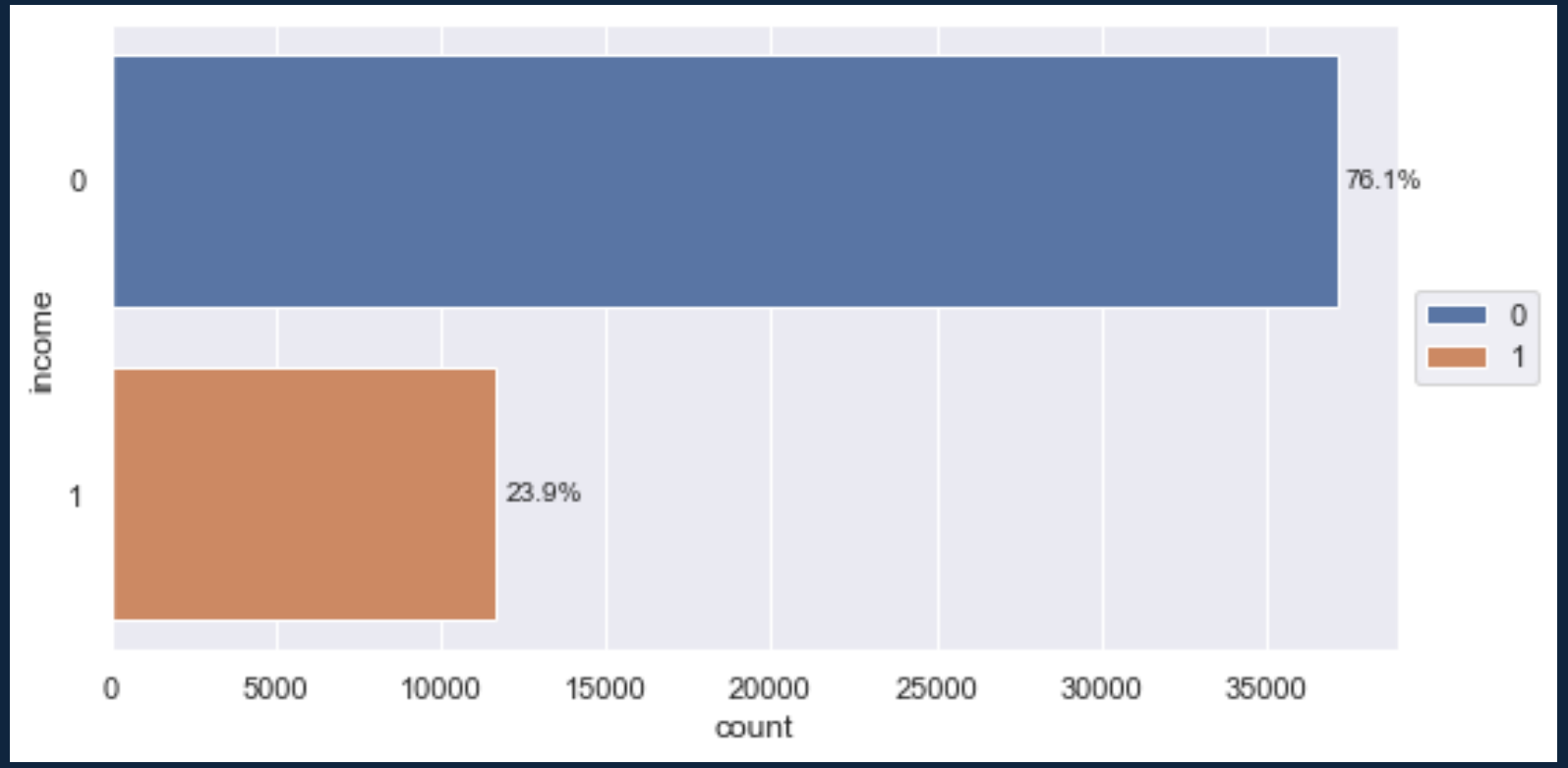
Transform Feature

Transformed the target feature 'income' where values are either >50k and <=50k to 0 when income is less than or equal 50,000 USD and 1 when income is greater than or equal to 50,000 USD

	race	sex	income
0	White	Male	<=50K
1	White	Male	<=50K
2	White	Male	<=50K
3	Black	Male	<=50K
4	Black	Female	<=50K
...	...	...	...
48837	White	Female	<=50K.
48838	Black	Male	<=50K.
48839	White	Male	<=50K.
48840	Asian-Pac-Islander	Male	<=50K.
48841	White	Male	>50K.

	race	sex	income
0	White	Male	0
1	White	Male	0
2	White	Male	0
3	Black	Male	0
4	Black	Female	0
...	...	...	...
48837	White	Female	0
48838	Black	Male	0
48839	White	Male	0
48840	Asian-Pac-Islander	Male	0
48841	White	Male	1

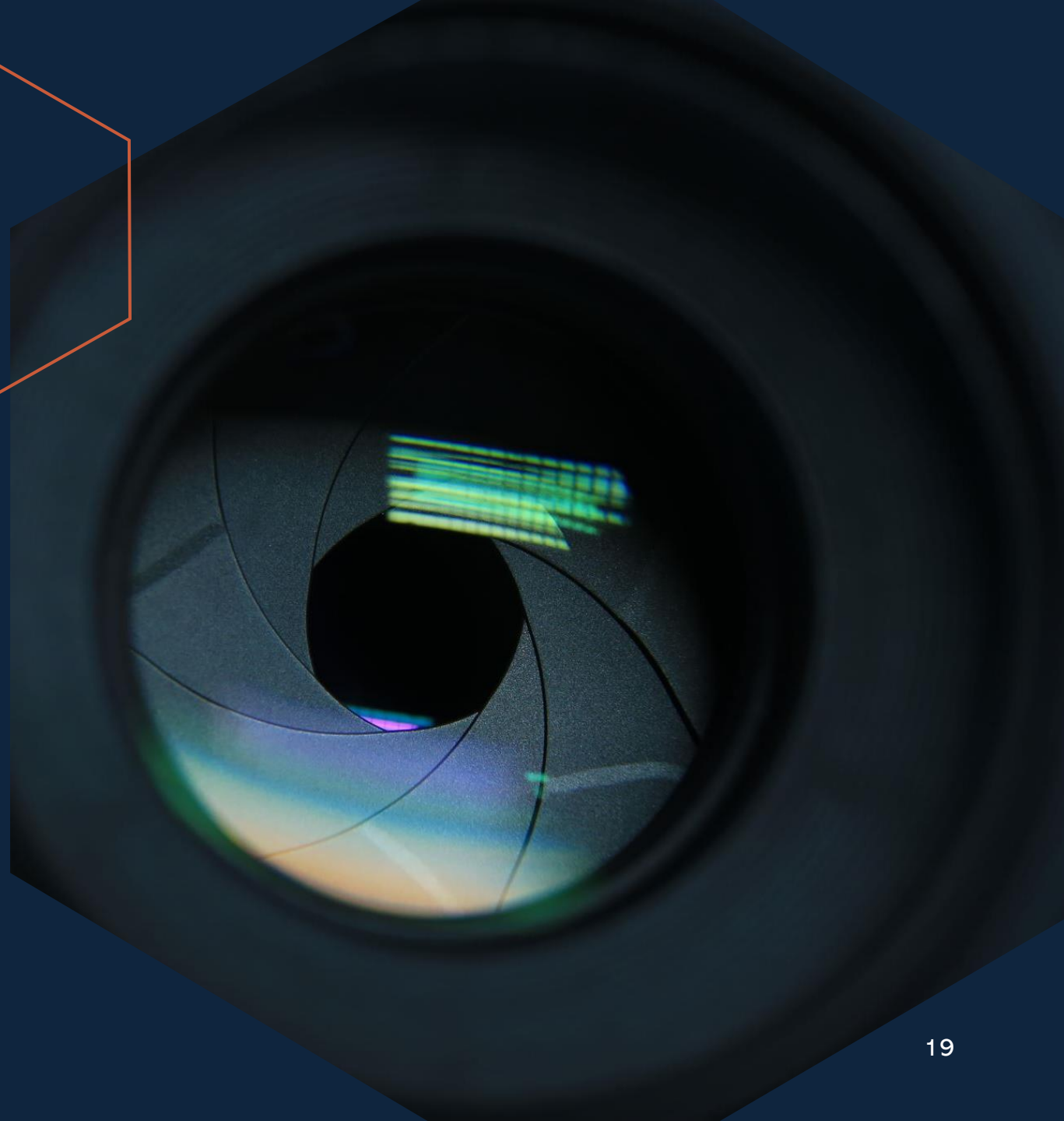
Count dominated by <=50k



With the total of 48,842 participants, the graph shows that 37,168 (76%) participants of the census dataset are income earners below or equal to 50,000 USD while 11,674 (24%) of the participants are earning greater than 50,000 USD.

Objective

To Determine the probability of a person to earn greater than 50,000 USD based on the features on the census dataset

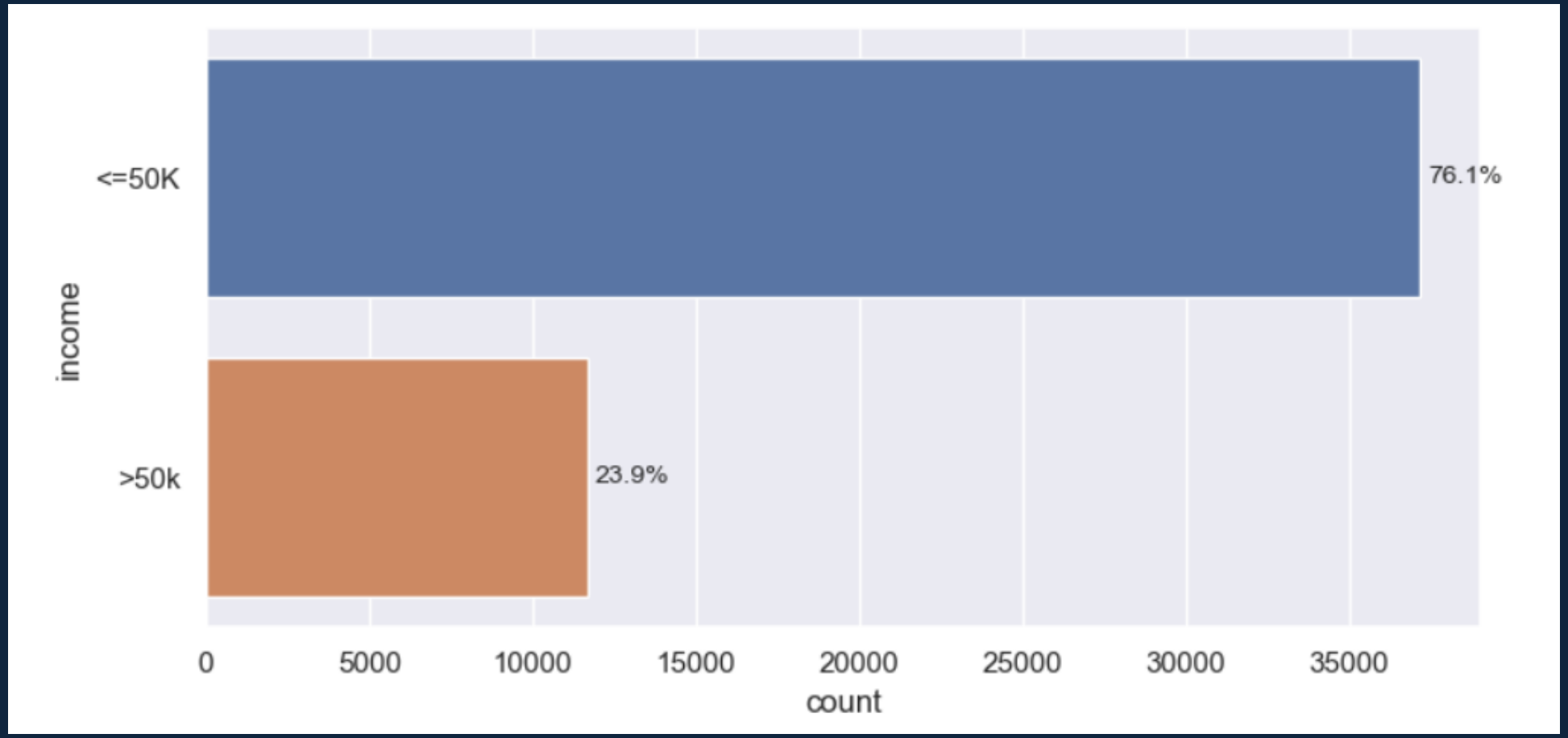


Exploratory Data Analysis

Target Feature and Methodology

Target Feature: Income >50k (affluence rate)

The base probability of having an income >50k is **24%.**



Univariate Analysis

Methodology

Cross tabulation of feature and income

income	0	1	All
sex			
Female	14423	1769	16192
Male	22732	9918	32650
All	37155	11687	48842

Predictive power of a feature vs the affluence rate:

$$\frac{P(Feature \cap > 50k)}{P(Feature)}$$

Methodology - Example

Cross tabulation of Sex and Income

income	0	1	All
sex			
Female	14423	1769	16192
Male	22732	9918	32650
All	37155	11687	48842

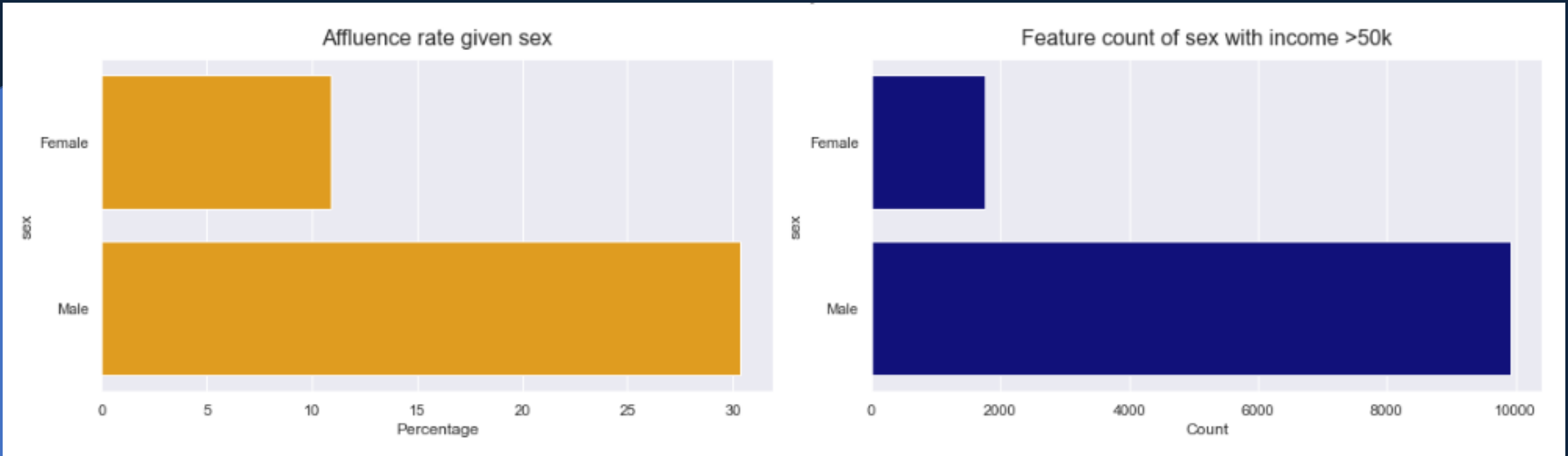
Predictive power of sex vs the affluence rate:

$$\frac{P(Male \cap >50k)}{P(All\ Male)} = \frac{9918}{32650}$$

$$\frac{P(Male \cap >50k)}{P(All\ Male)} = 30.4\%$$

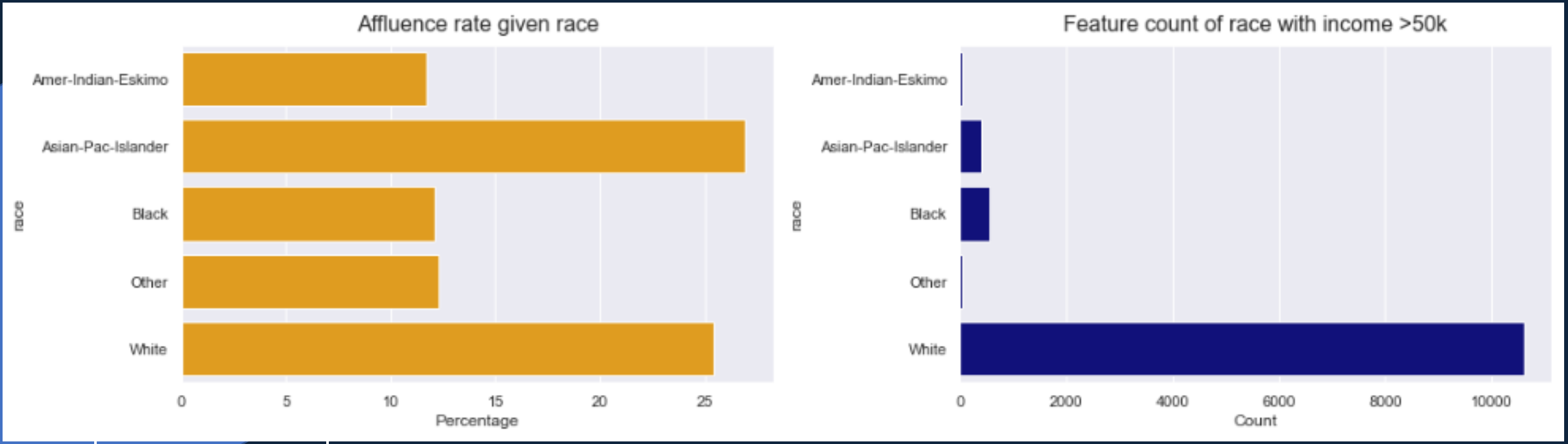
Univariate Analysis: Sex

Males have higher chance than *females*.



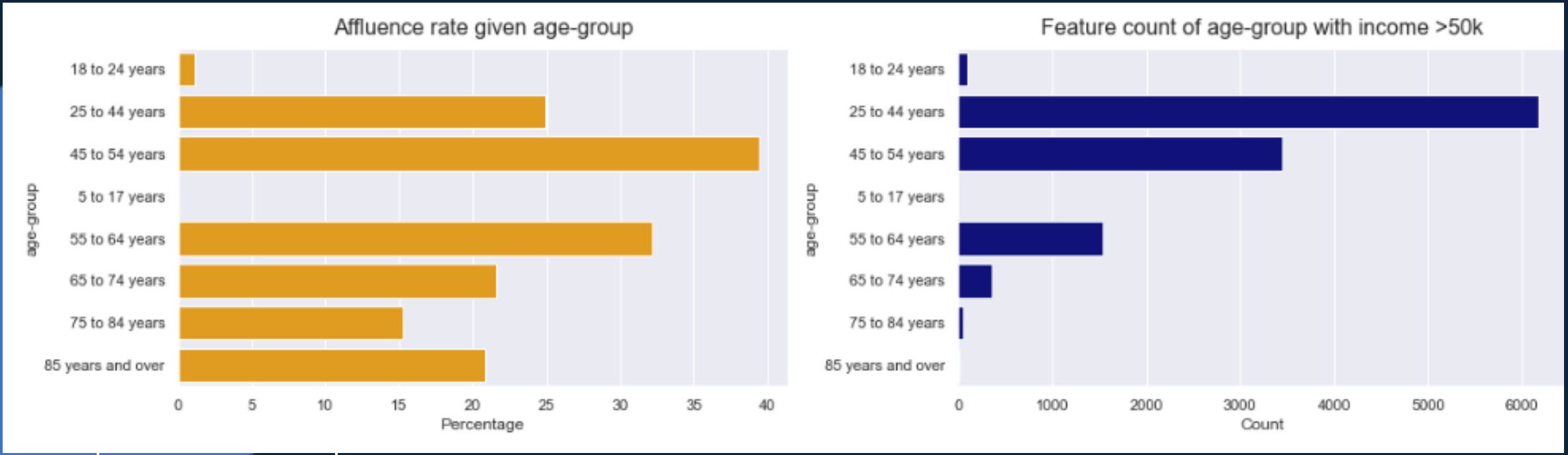
Univariate Analysis: Race

Asians and *white* people have an advantage.



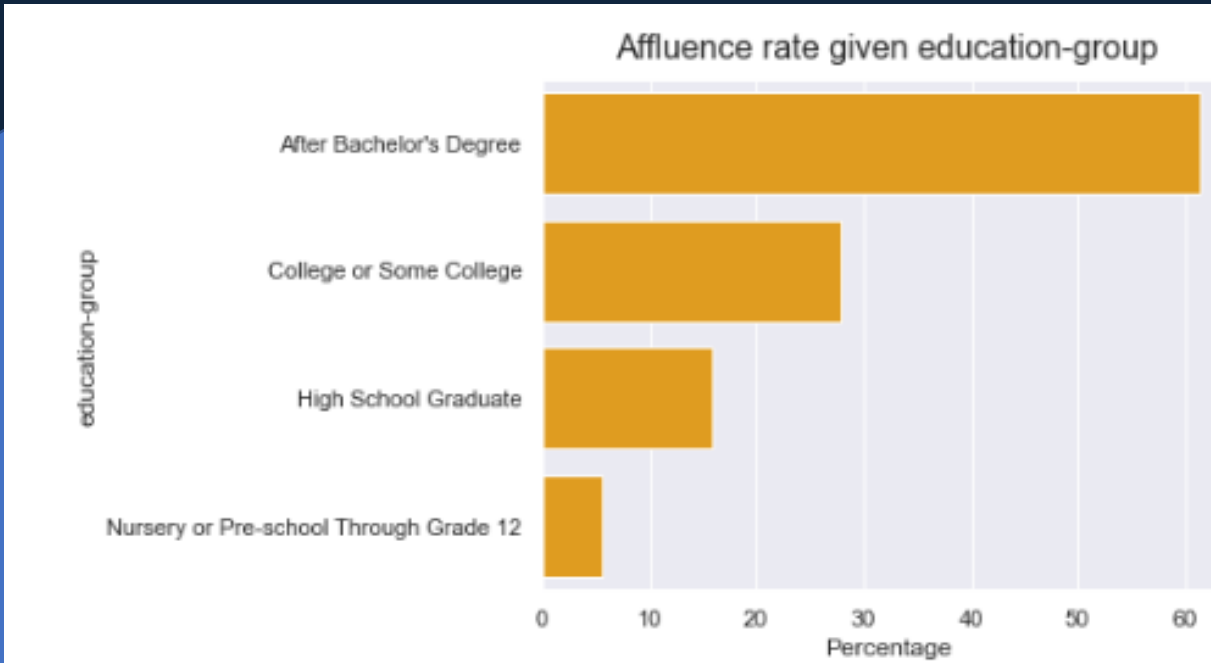
Univariate Analysis: Age

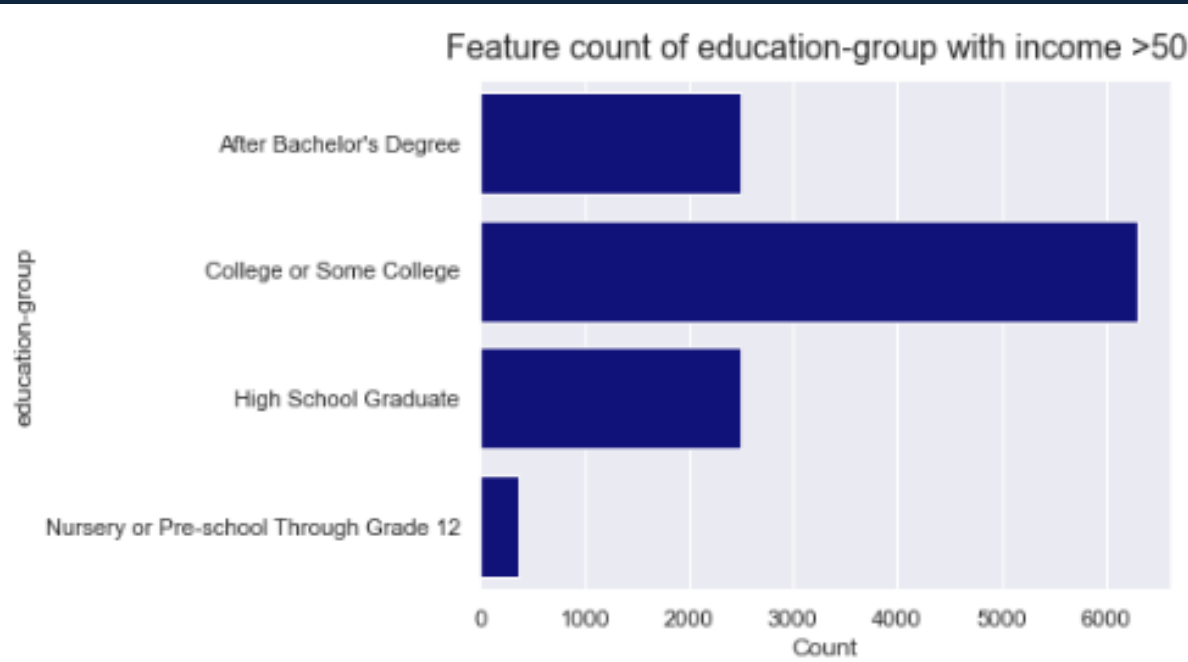
Age group between *45 to 54 years old* is more likely to have an income >50k.



Univariate Analysis: Education

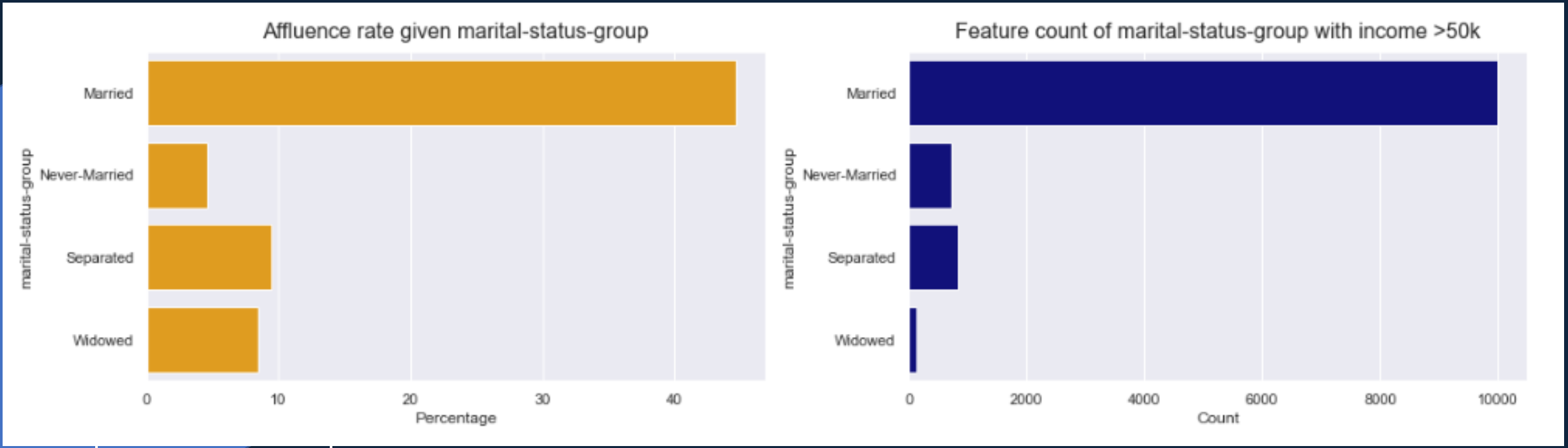
Having post-graduate degrees (*after bachelor's degree*) increases the probability by a wide margin.





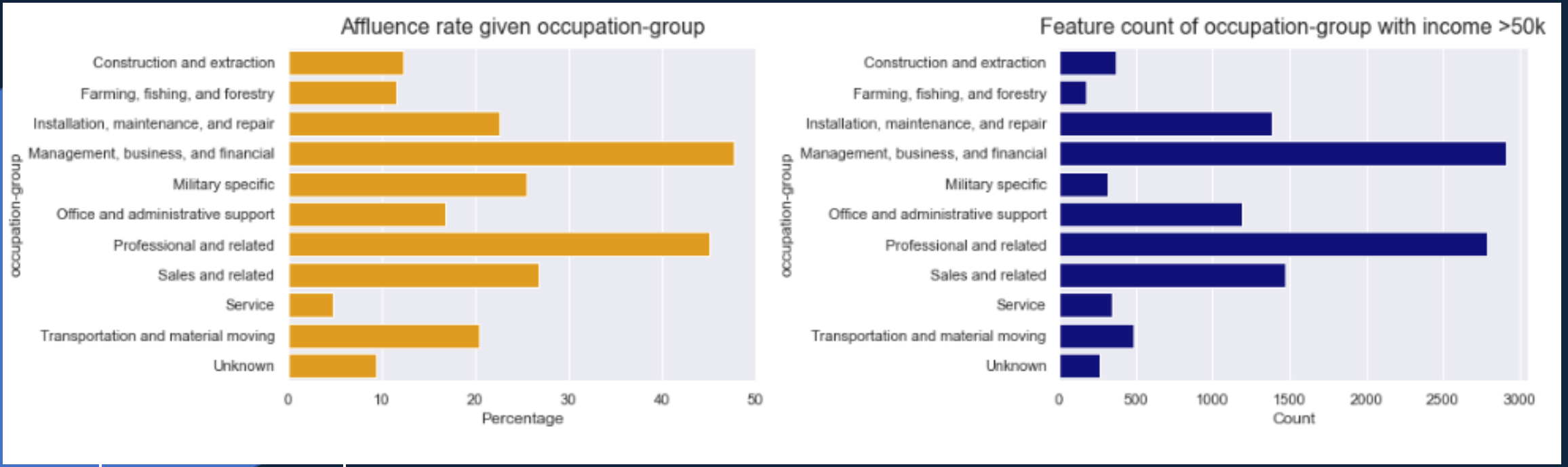
Univariate Analysis: Marital Status

Married people are more likely to have an income >50k.



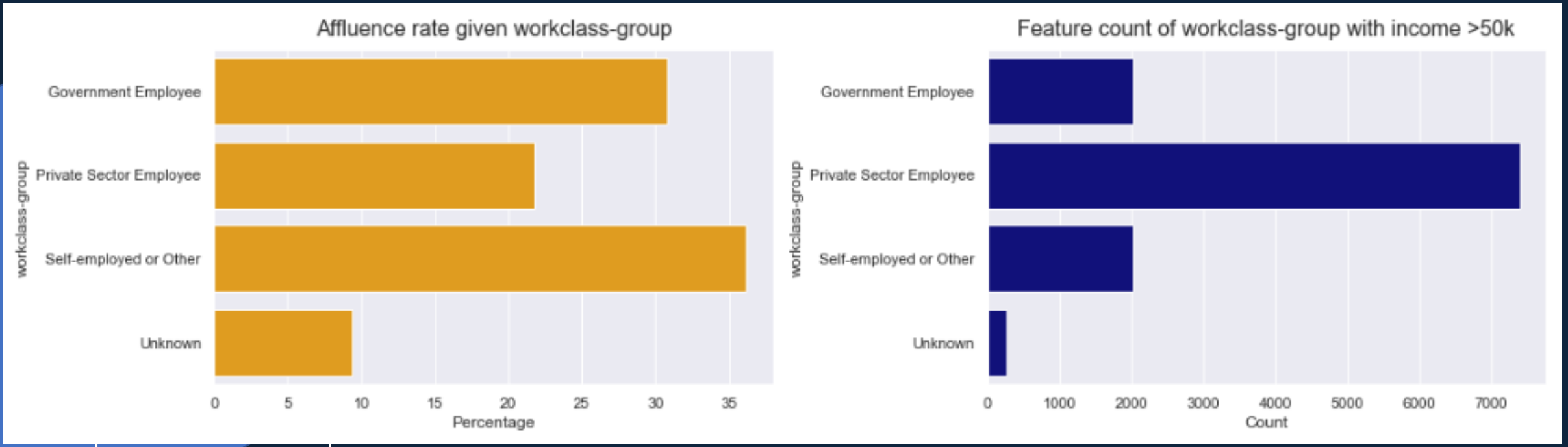
Univariate Analysis: Occupation

Management, business, finance, and *Professional and related* are the top-paying jobs.



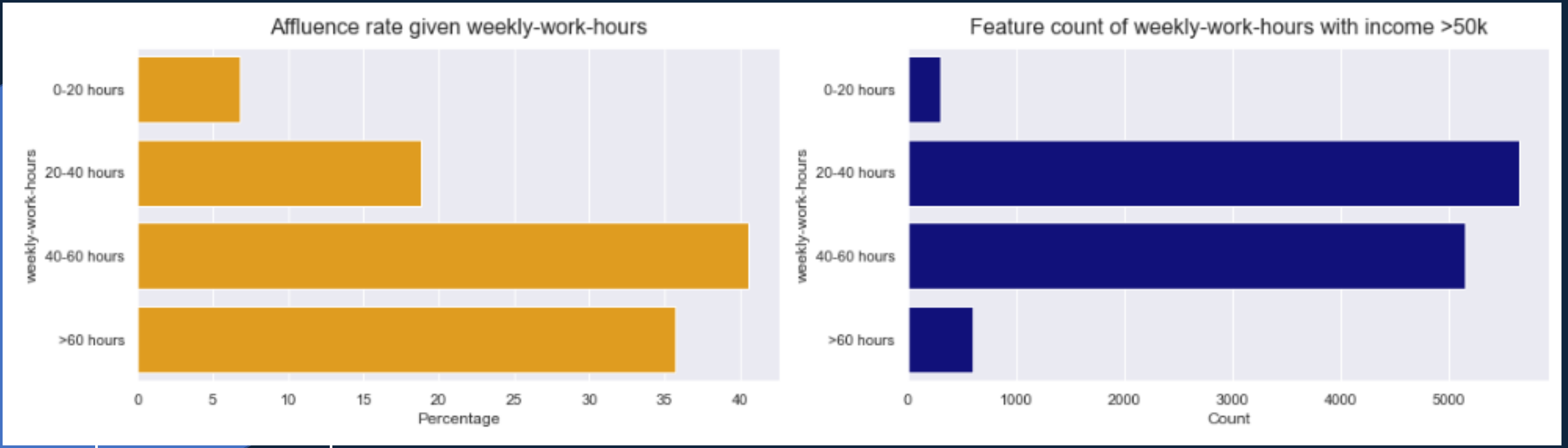
Univariate Analysis: Work Class

Being *self-employed* gives the highest likelihood of having an income >50k.



Univariate Analysis: Weekly Work Hours

Working more than *60 hours* has diminishing returns.



Top five features for affluence rate

1. **Education**
    - 61% probability for *post-graduate degree holders*.
2. **Occupation**
    - 48% for individuals working in *management, business, and finance*
    - 45% for *professionals*
3. **Marital Status**
    - 45% for *married* people
4. **Weekly Work Hours**
    - 41% for *40-60 hours* of work weekly
5. **Age**
    - 40% for *45 to 54 years old* age group

Bivariate Analysis

Methodology

Cross tabulation of feature and income filtered from another feature

income	0	1	All
sex			
Female	1384	1121	2505
Male	11034	8877	19911
All	12418	9998	22416

Predictive power of a feature vs the affluence rate:

$$\frac{P(Feature \cap > 50k)}{P(Feature)}$$

Methodology - Example

Cross tabulation of sex and income filtered for married individuals

income	0	1	All
sex			
Female	1384	1121	2505
Male	11034	8877	19911
All	12418	9998	22416

Predictive power of a feature vs the affluence rate:

$$\frac{P(Male \cap >50k)}{P(All\ Male)} = \frac{8877}{19911}$$

$$\frac{P(Male \cap >50k)}{P(All\ Male)} = 44.6\%$$

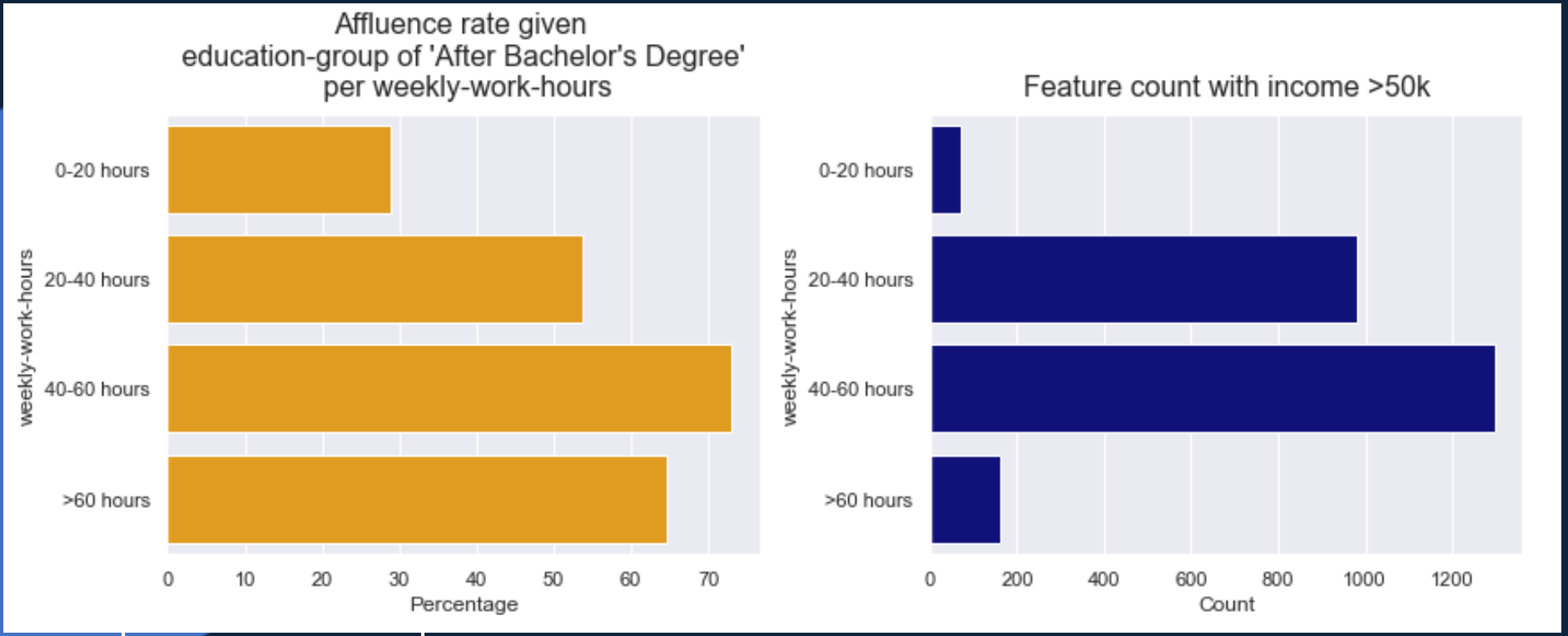
Education vs Work Hours

- After Bachelor's Degree
- College or Some College
- High School
- Nursery or Pre-school through Grade 12

- 0 - 20 hours
- 20 – 40 hours
- 40 to 60 hours
- >60 hours

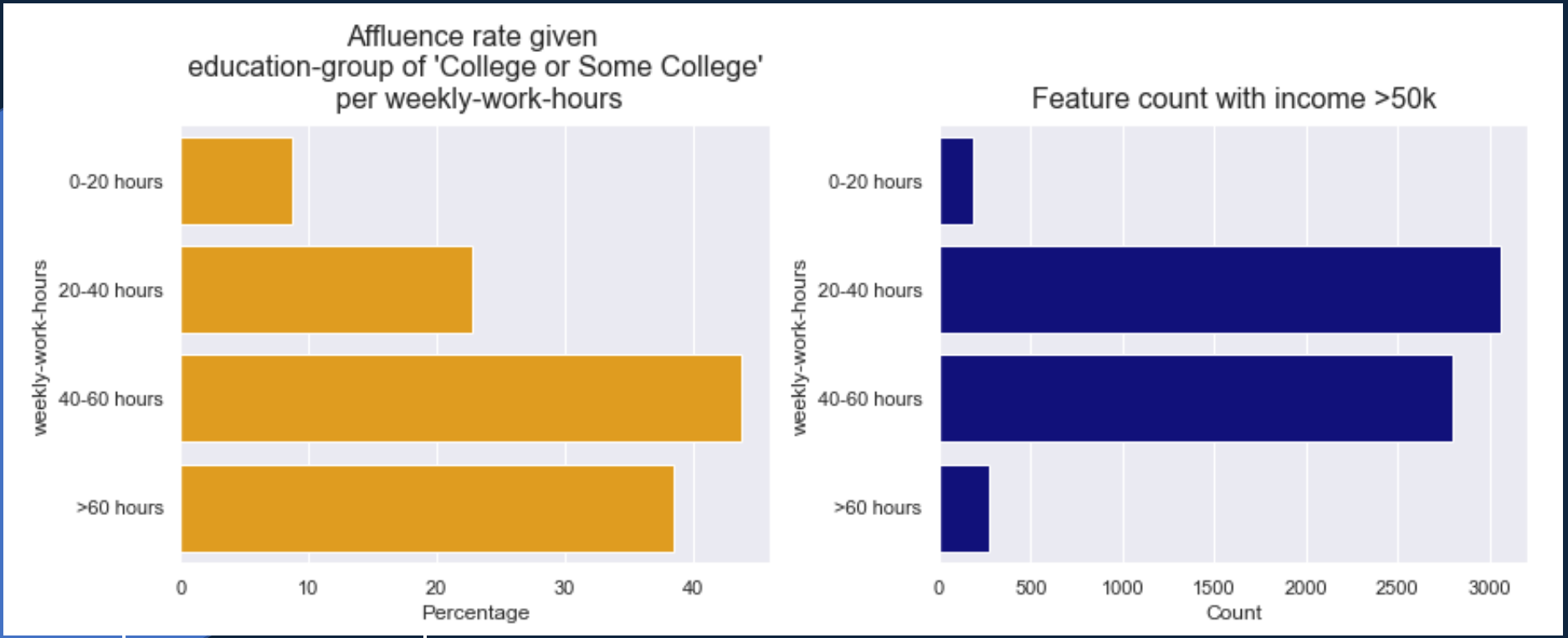
Bivariate Analysis: Education vs Work Hours

Working for *60 hours* with *post-graduate degrees* is highly favored.



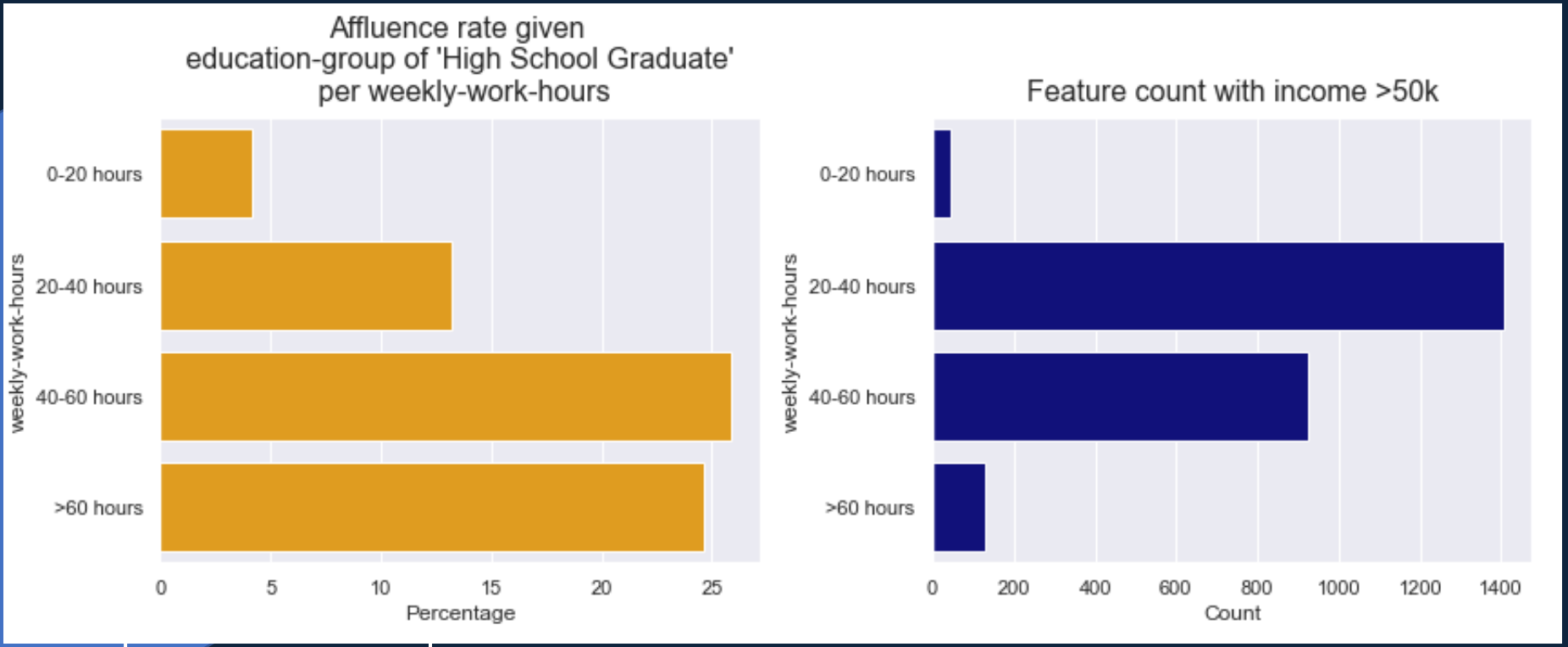
Bivariate Analysis: Education vs Work Hours

Working for *60 hours* with *post-graduate degrees* is highly favored.



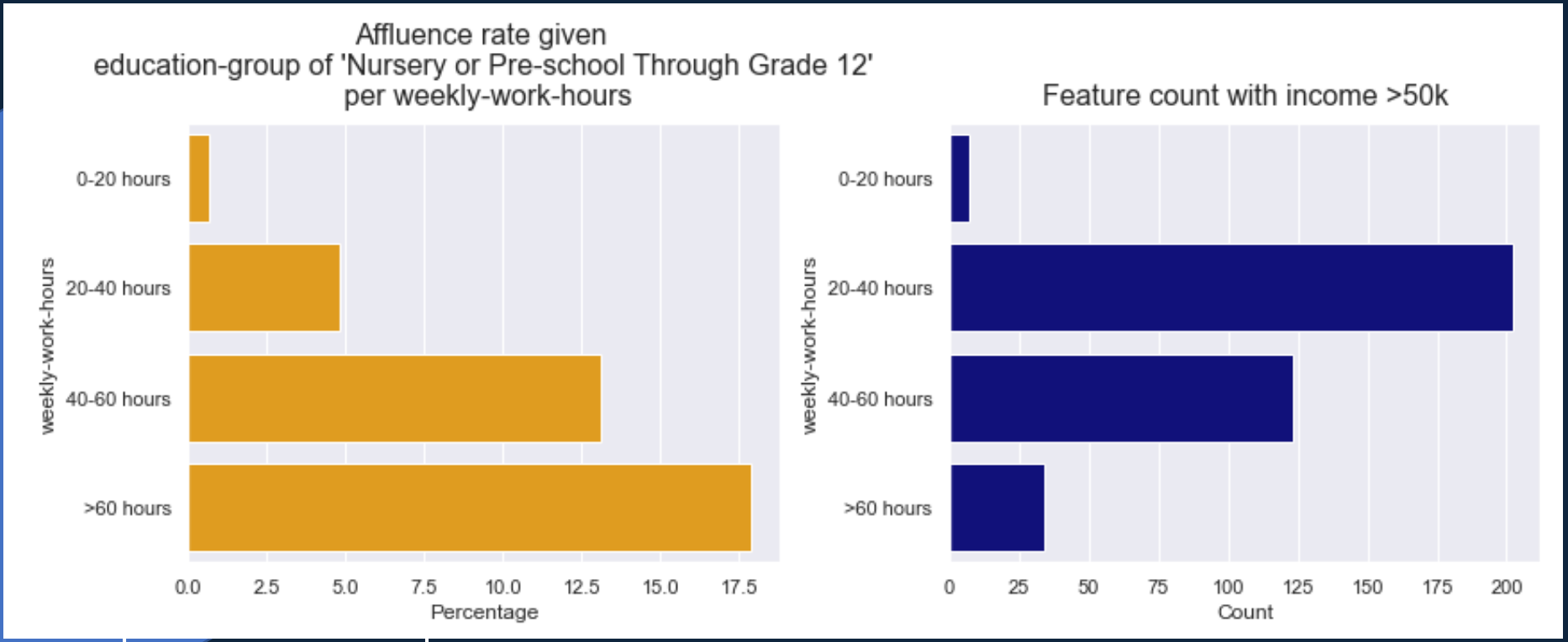
Bivariate Analysis: Education vs Work Hours

Working for *60 hours* with *post-graduate degrees* is highly favored.



Bivariate Analysis: Education vs Work Hours

Working for *60 hours* with *post-graduate degrees* is highly favored.



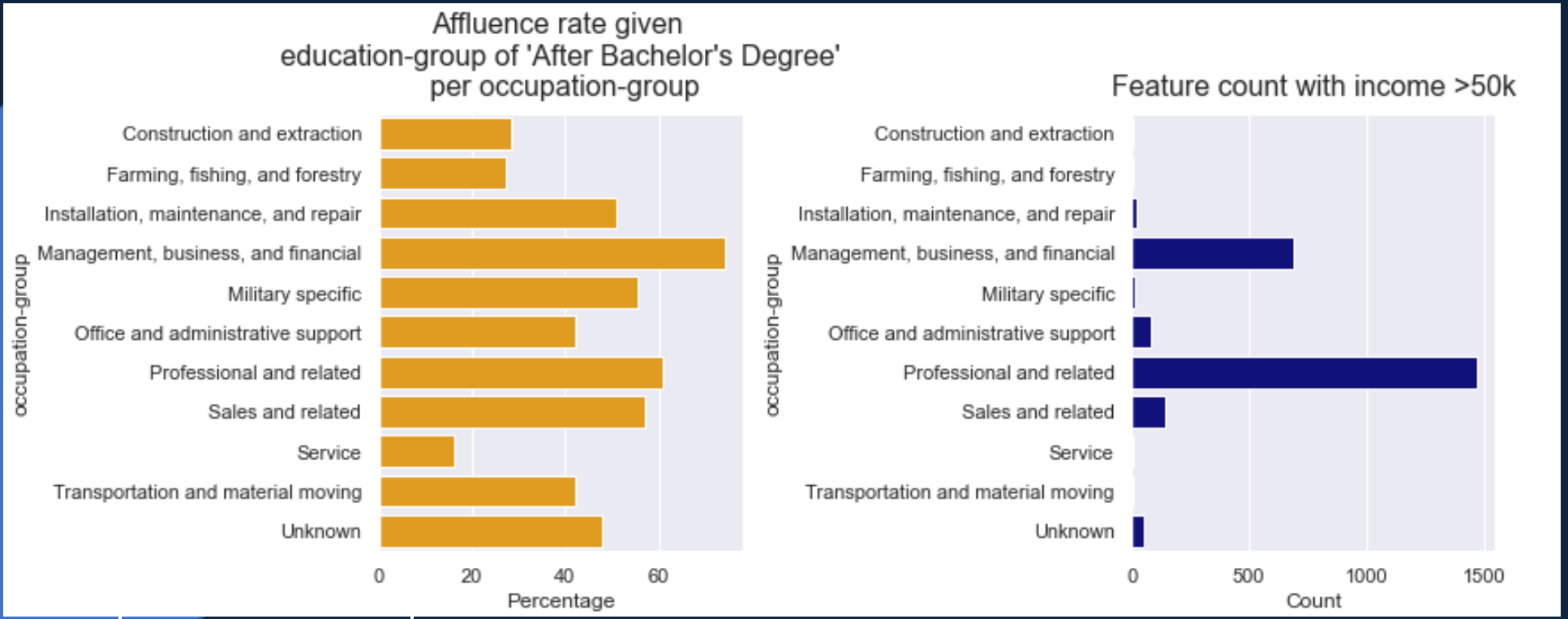
Education vs Occupation

- After Bachelor's Degree
- College or Some College
- High School
- Nursery or Pre-school through Grade 12

- Construction and Extraction
- Farming, fishing and forestry
- Installation, maintenance, and repair
- Management, business, and financial
- Military Specific
- Office and administrative support
- Professional and related
- Sales and related
- Service
- Transportation and material moving
- Unknown

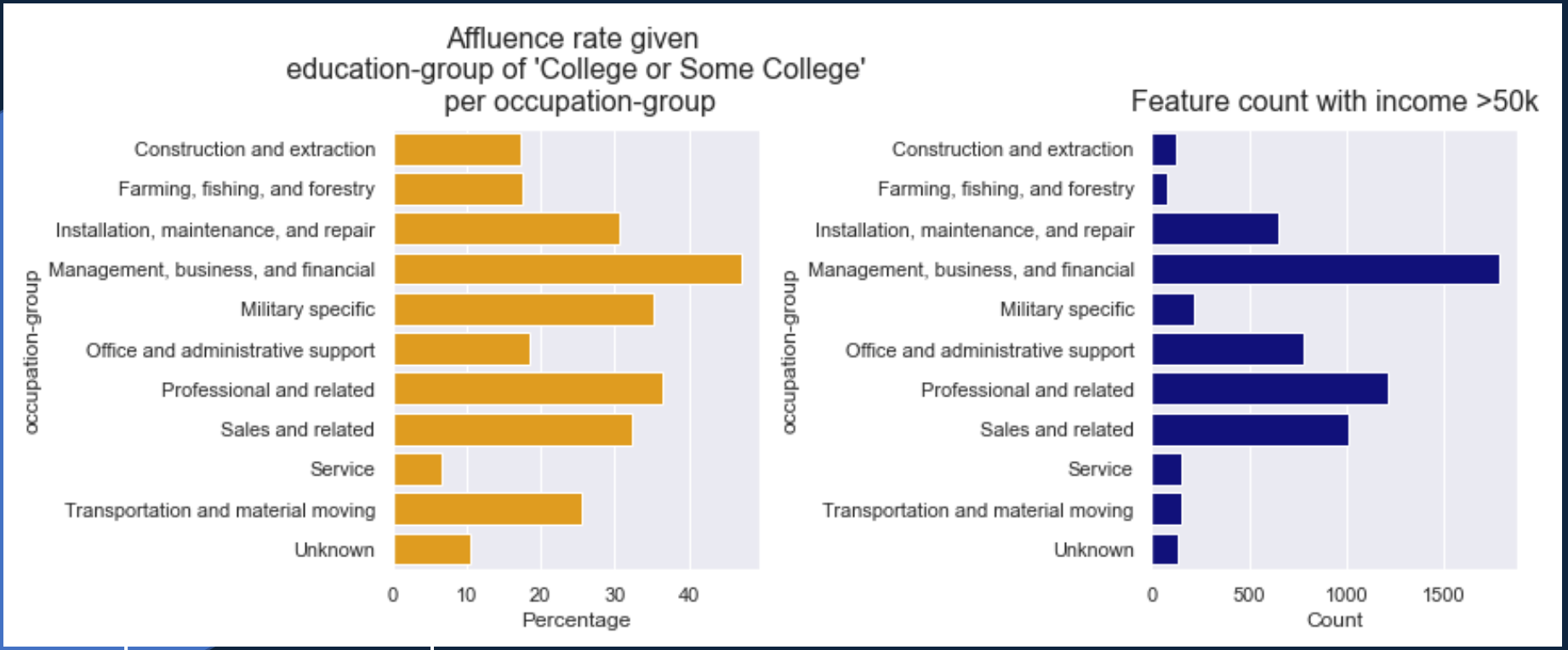
Bivariate Analysis: Education vs Occupation

Management, professionals, and *sales* are the top occupations for bachelor's and post-graduate degree holders.



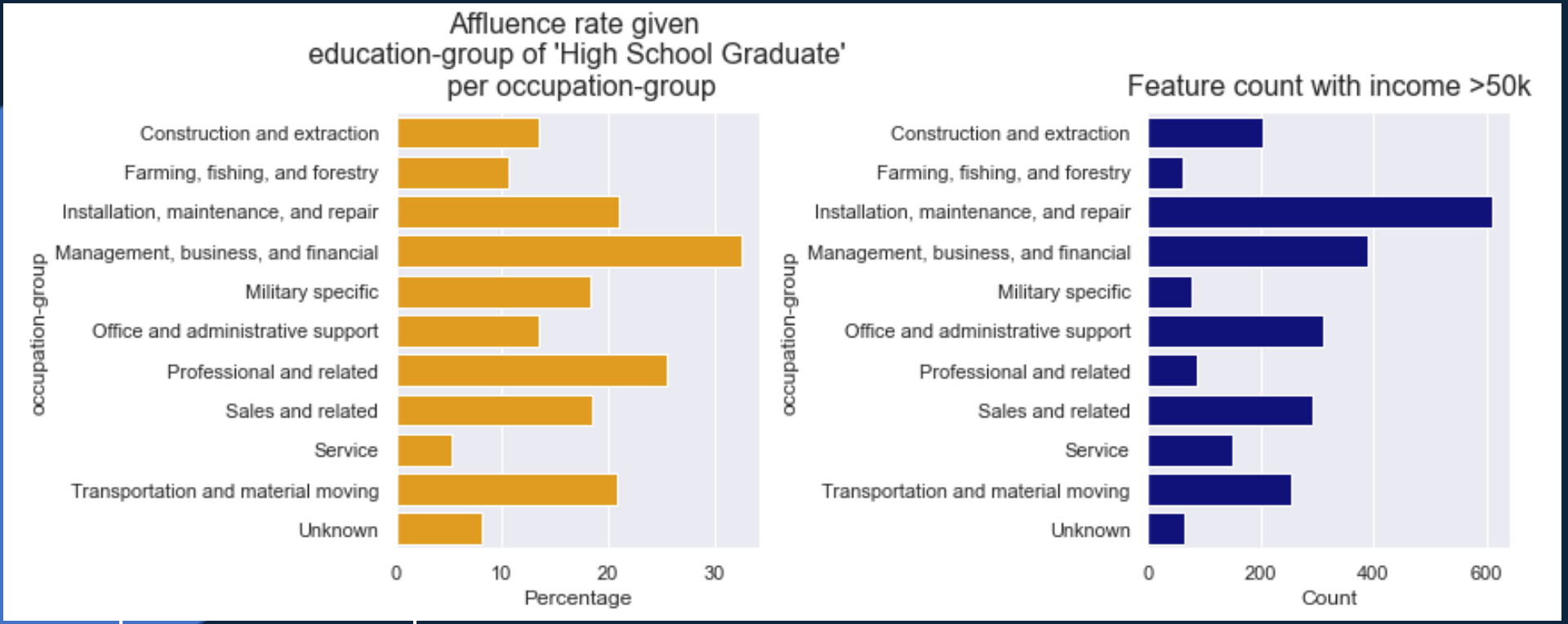
Bivariate Analysis: Education vs Occupation

Management, professionals, and *sales* are the top occupations for bachelor's and post-graduate degree holders.



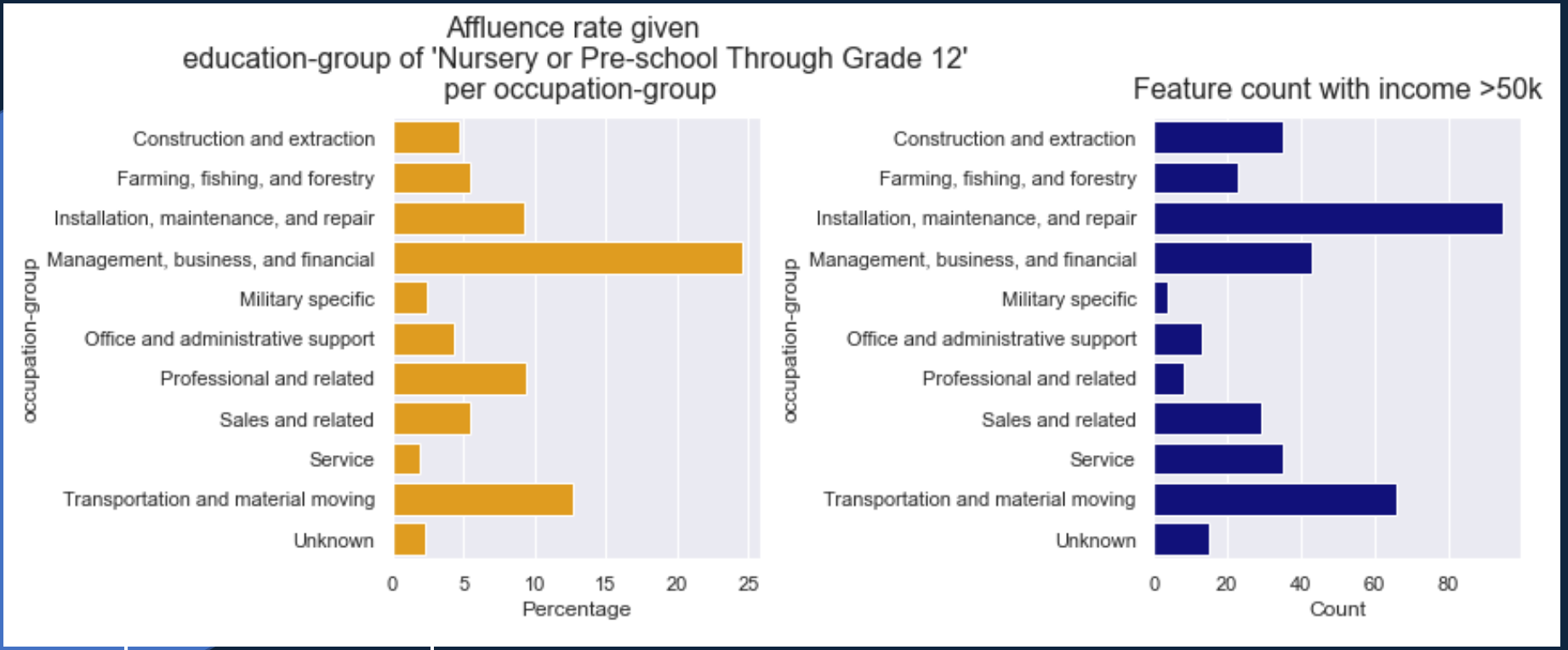
Bivariate Analysis: Education vs Occupation

Installation, maintenance, repair, and *transportation and material moving* ranks second and third for non-degree holders.



Bivariate Analysis: Education vs Occupation

Installation, maintenance, repair, and *transportation and material moving* rank second and third for non-degree holders.

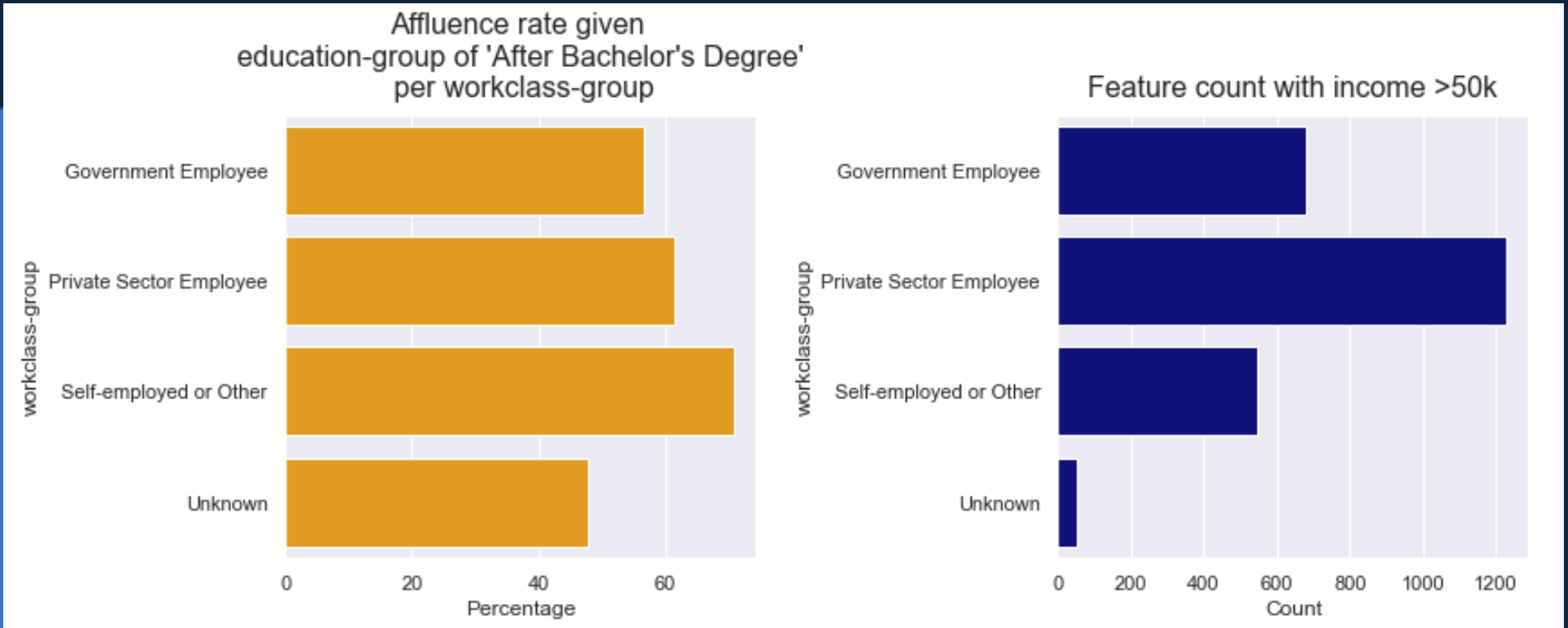


Education vs Work Class

- After Bachelor's Degree
- College or Some College
- High School
- Nursery or Pre-school through Grade 12
- Government Employee
- Private Sector Employee
- Self-employed or Other
- Unknown

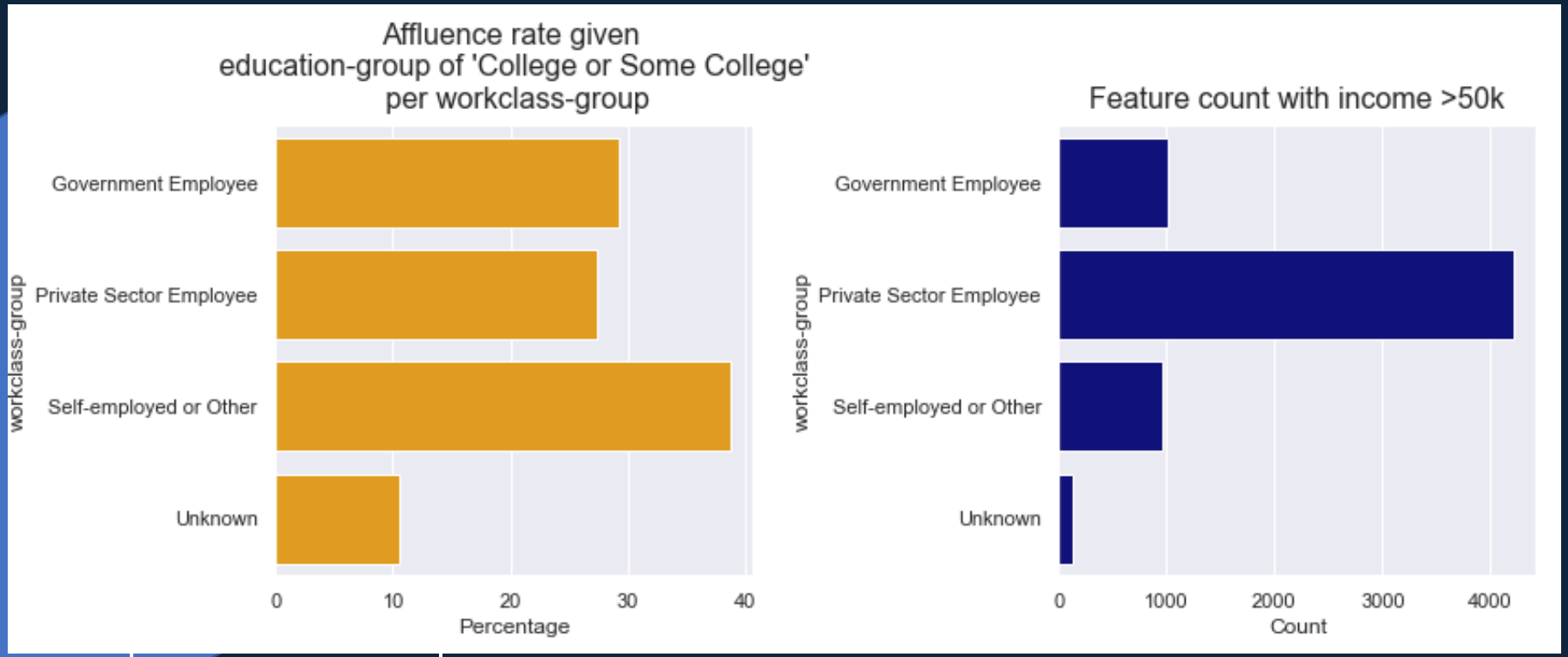
Bivariate Analysis: Education vs Work Class

Being *self-employed* gives the highest likelihood of having an income >50k among all education groups.



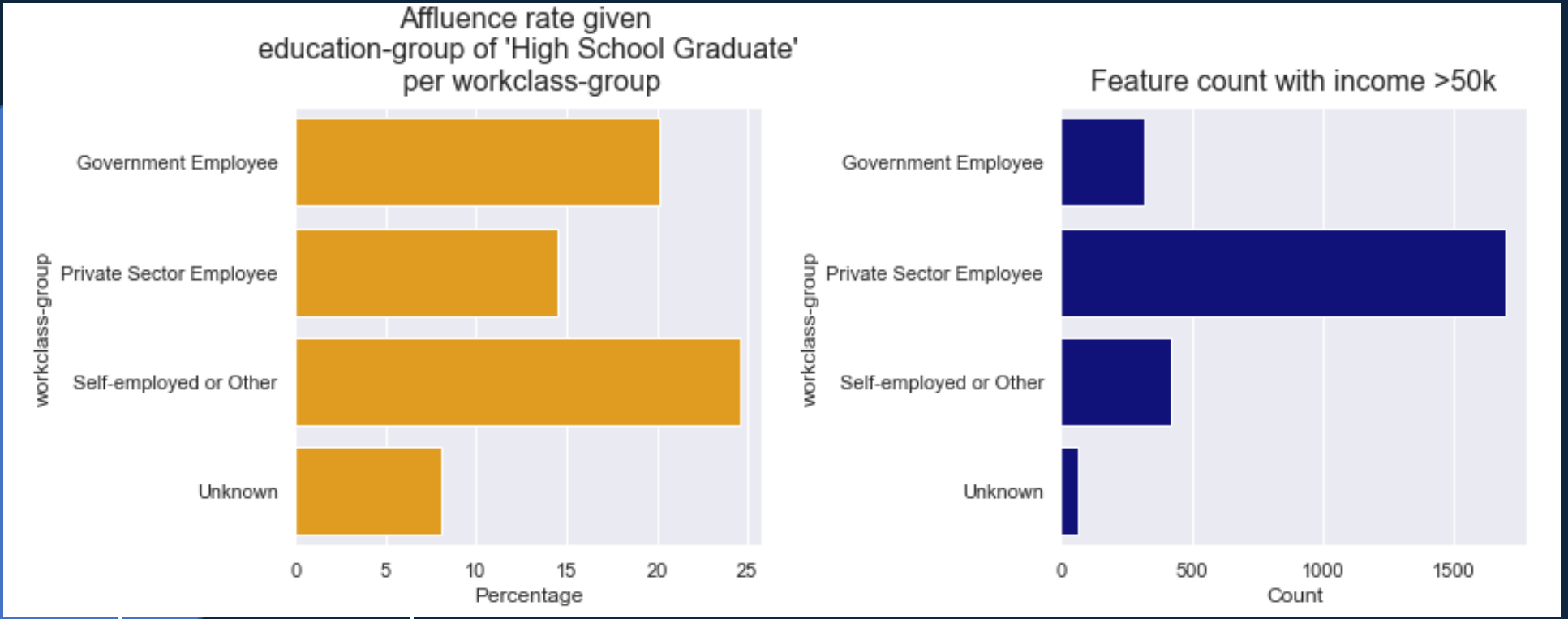
Bivariate Analysis: Education vs Work Class

Being *self-employed* gives the highest likelihood of having an income >50k among all education groups.



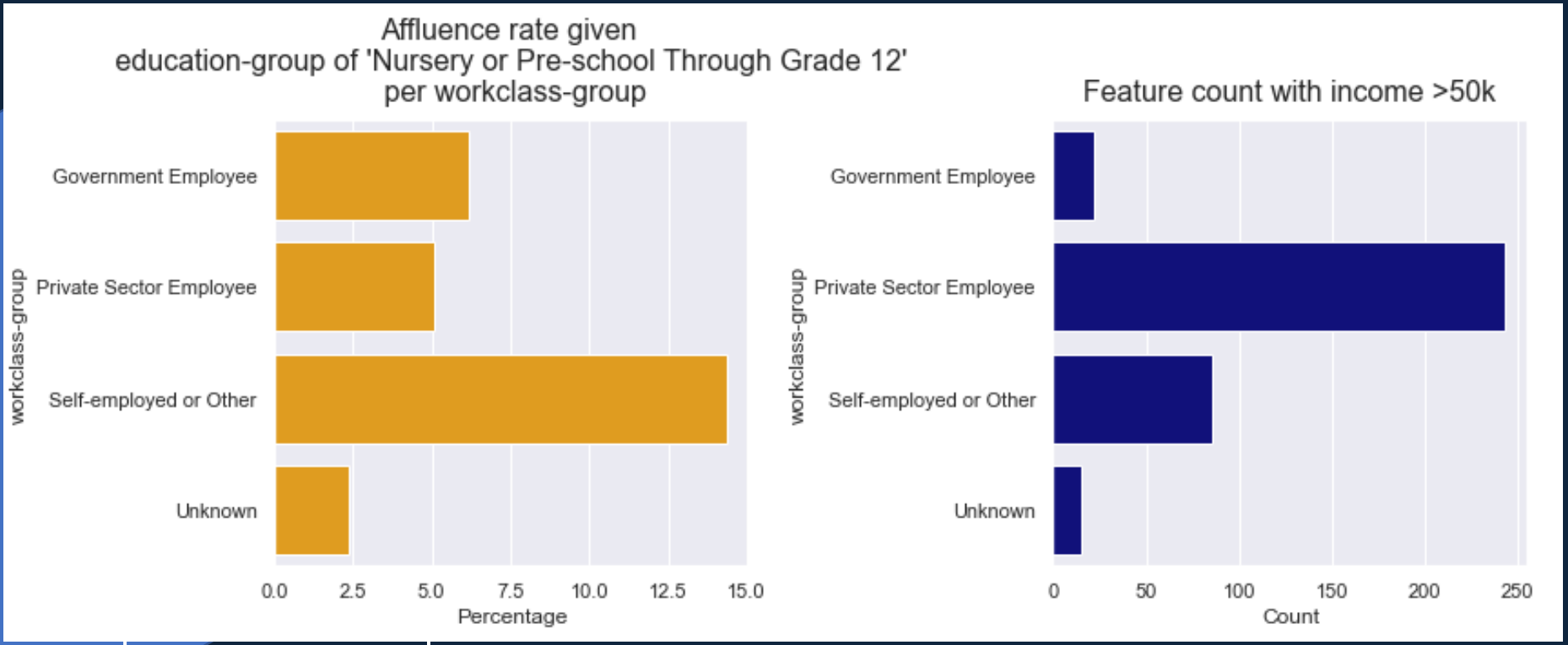
Bivariate Analysis: Education vs Work Class

Being *self-employed* gives the highest likelihood of having an income >50k among all education groups.



Bivariate Analysis: Education vs Work Class

Being *self-employed* gives the highest likelihood of having an income >50k among all education groups.



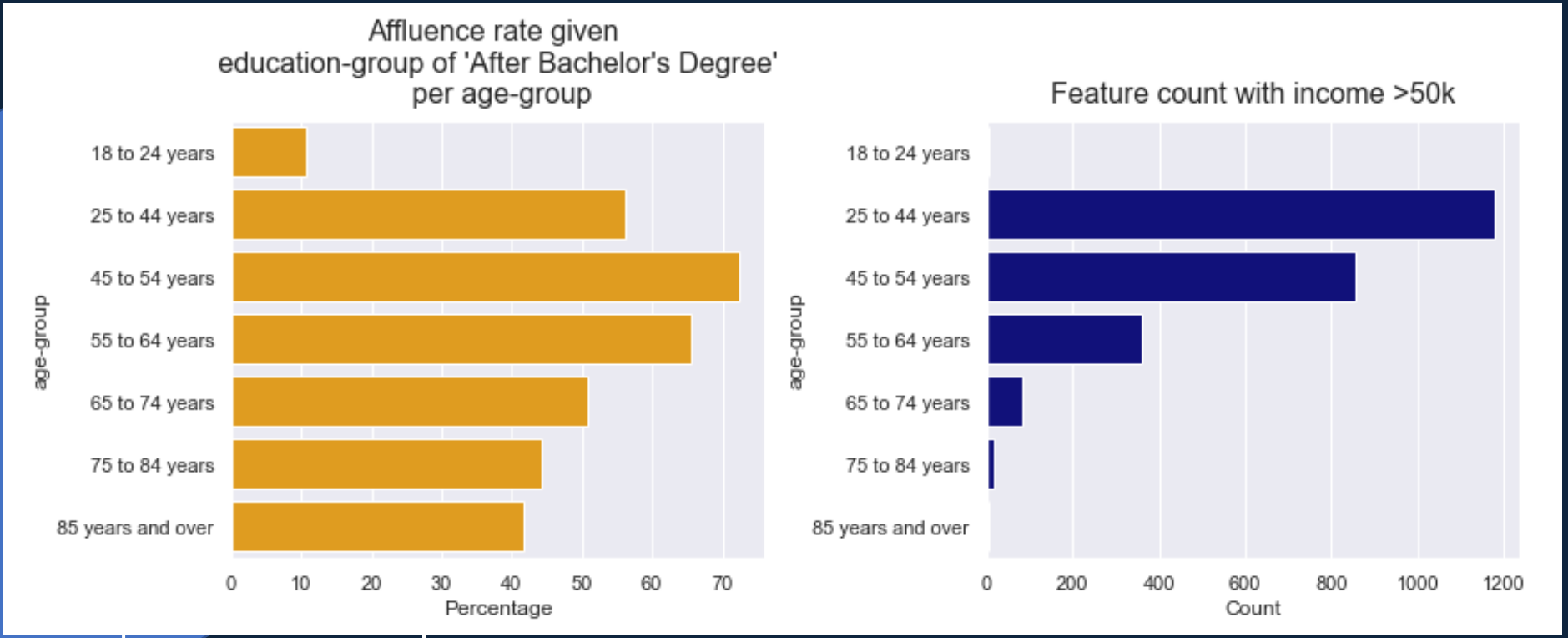
Education vs Age

- After Bachelor's Degree
- College or Some College
- High School
- Nursery or Pre-school through Grade 12

- 5 to 17 years
- 18 to 24 years
- 25 to 44 years
- 44 to 54 years
- 55 to 64 years
- 65 to 74 years
- 75 to 84 years
- 85 years and over

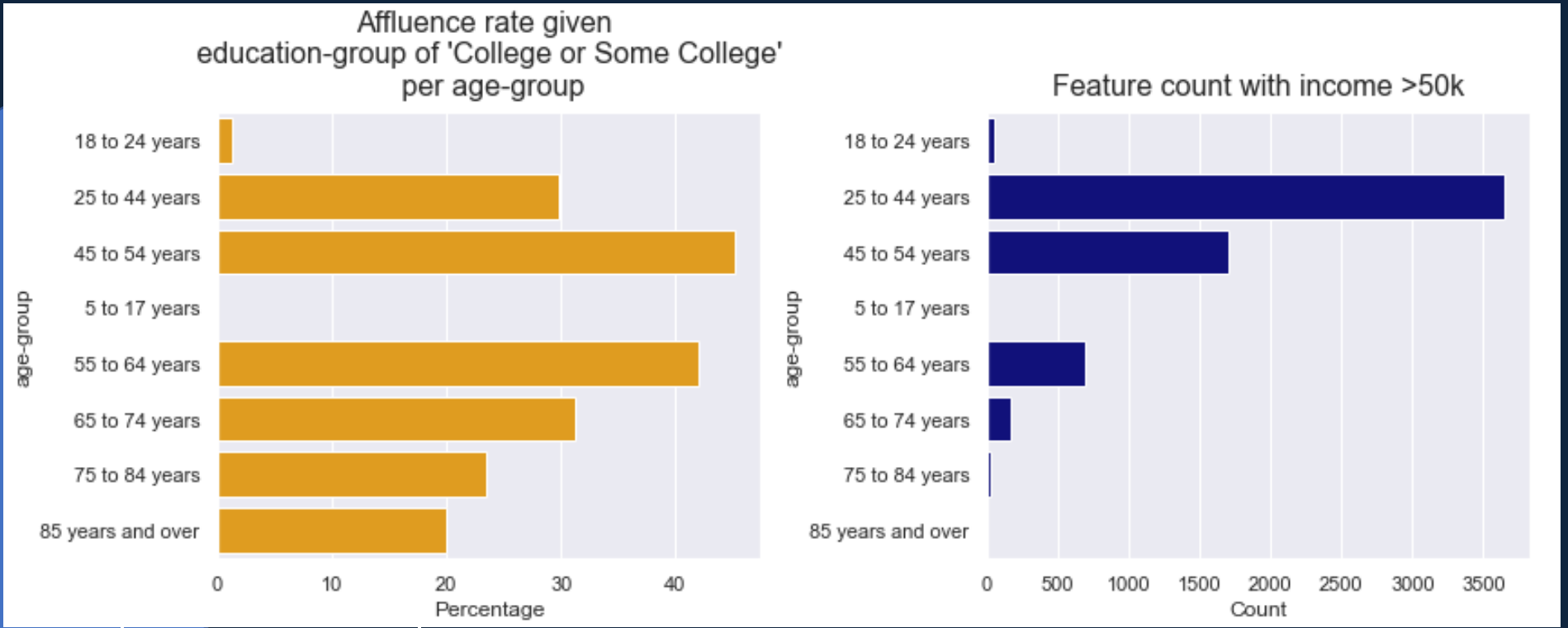
Bivariate Analysis: Education vs Age

Age groups between *45 to 54 years* and *55 to 64 years* that are degree holders are more likely to have an income >50k.



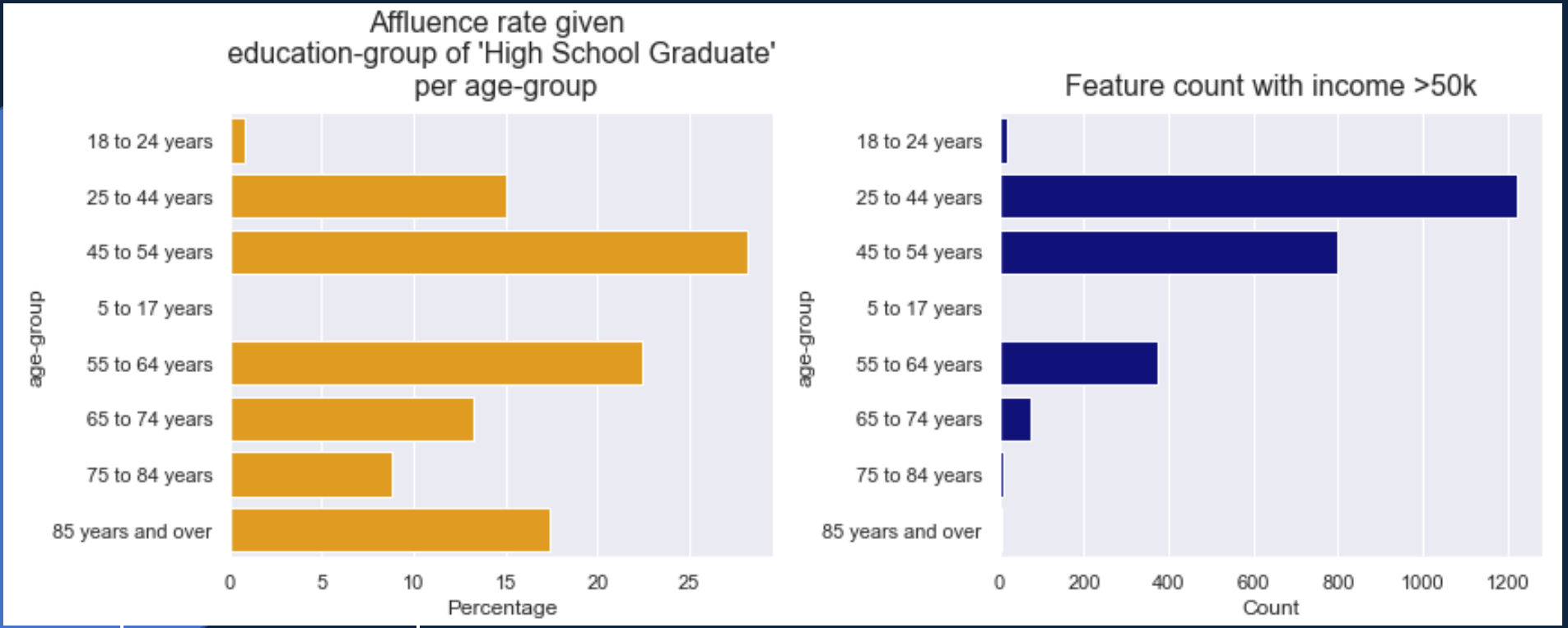
Bivariate Analysis: Education vs Age

Age groups between *45 to 54 years* and *55 to 64 years* that are degree holders are more likely to have an income >50k.



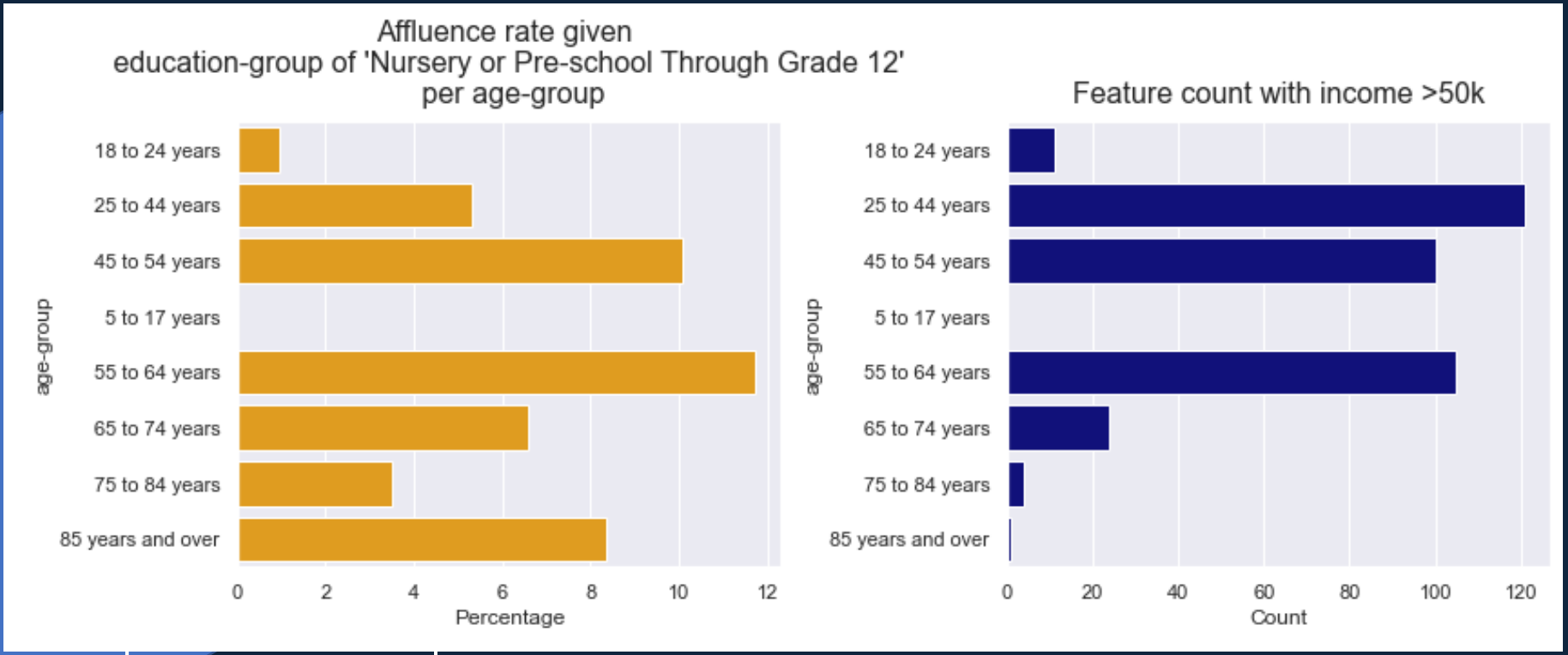
Bivariate Analysis: Education vs Age

Age groups between *45 to 54 years* and *55 to 64 years* that are degree holders are more likely to have an income >50k.



Bivariate Analysis: Education vs Age

Age groups between *45 to 54 years* and *55 to 64 years* that are degree holders are more likely to have an income >50k.



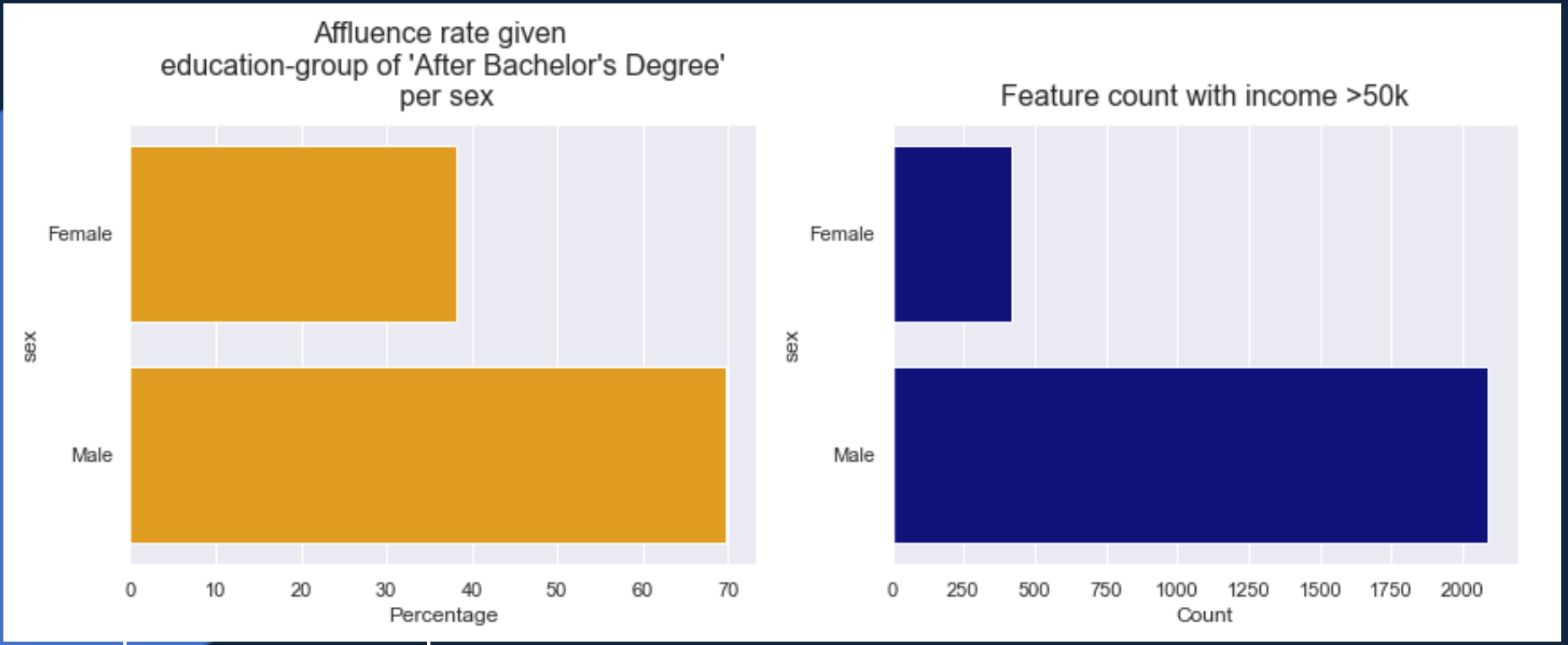
Education vs Sex

- After Bachelor's Degree
- College or Some College
- High School
- Nursery or Pre-school through Grade 12

- Female
- Male

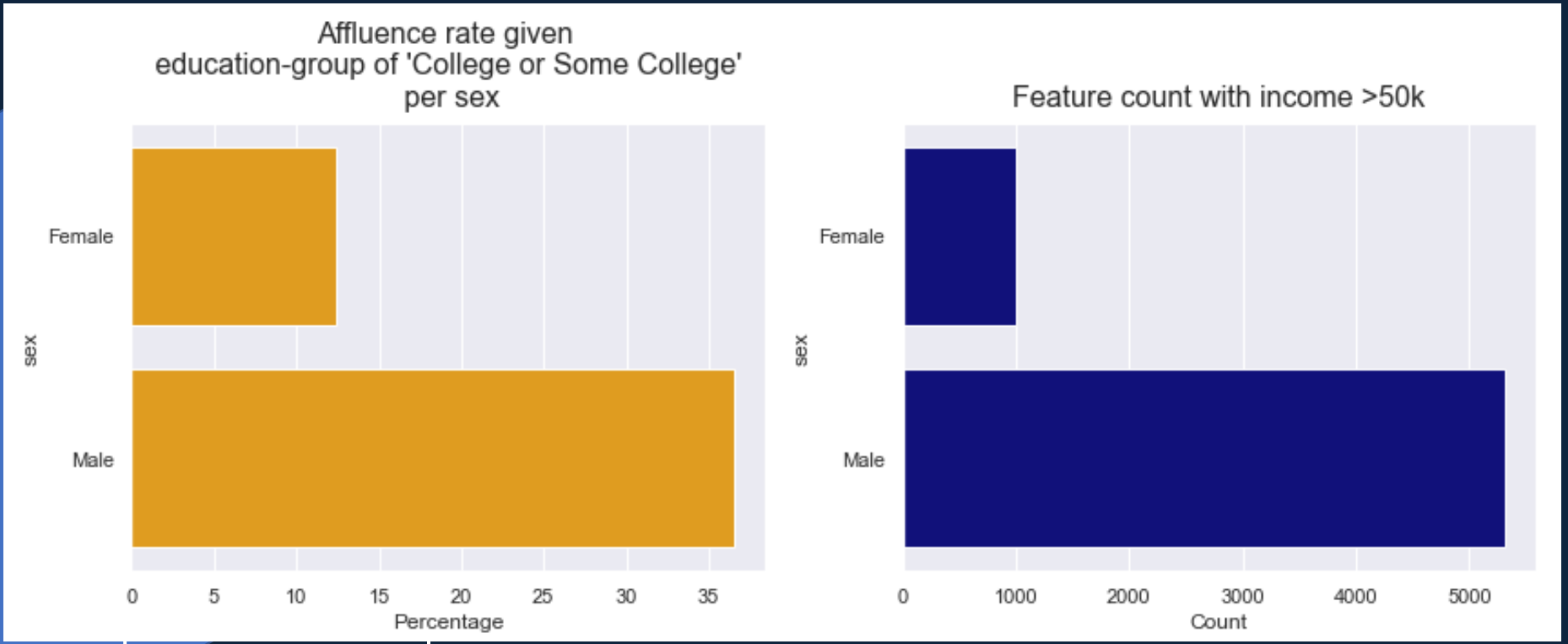
Bivariate Analysis: Education vs Sex

Males are more likely to have an income >50k than *females, g*iven the same educational attainment.



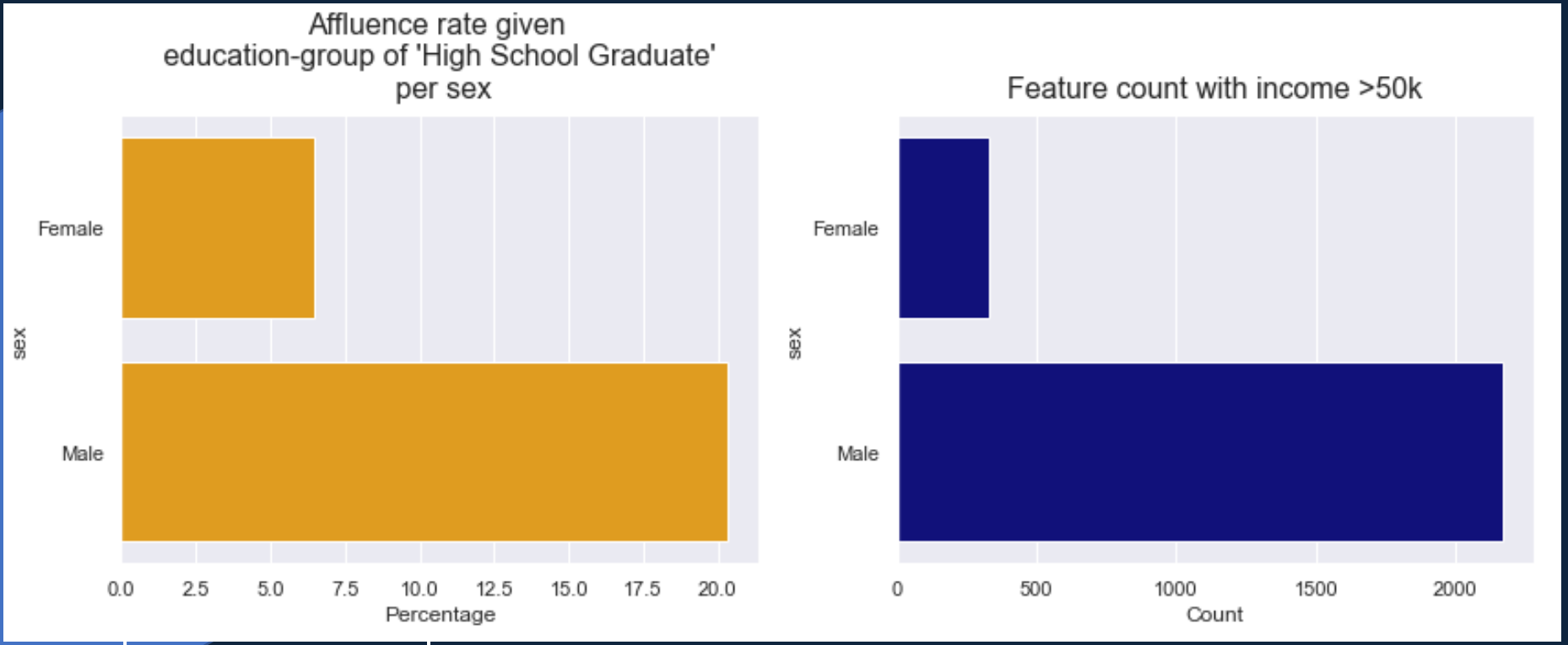
Bivariate Analysis: Education vs Sex

Males are more likely to have an income >50k than *females, g*iven the same educational attainment.



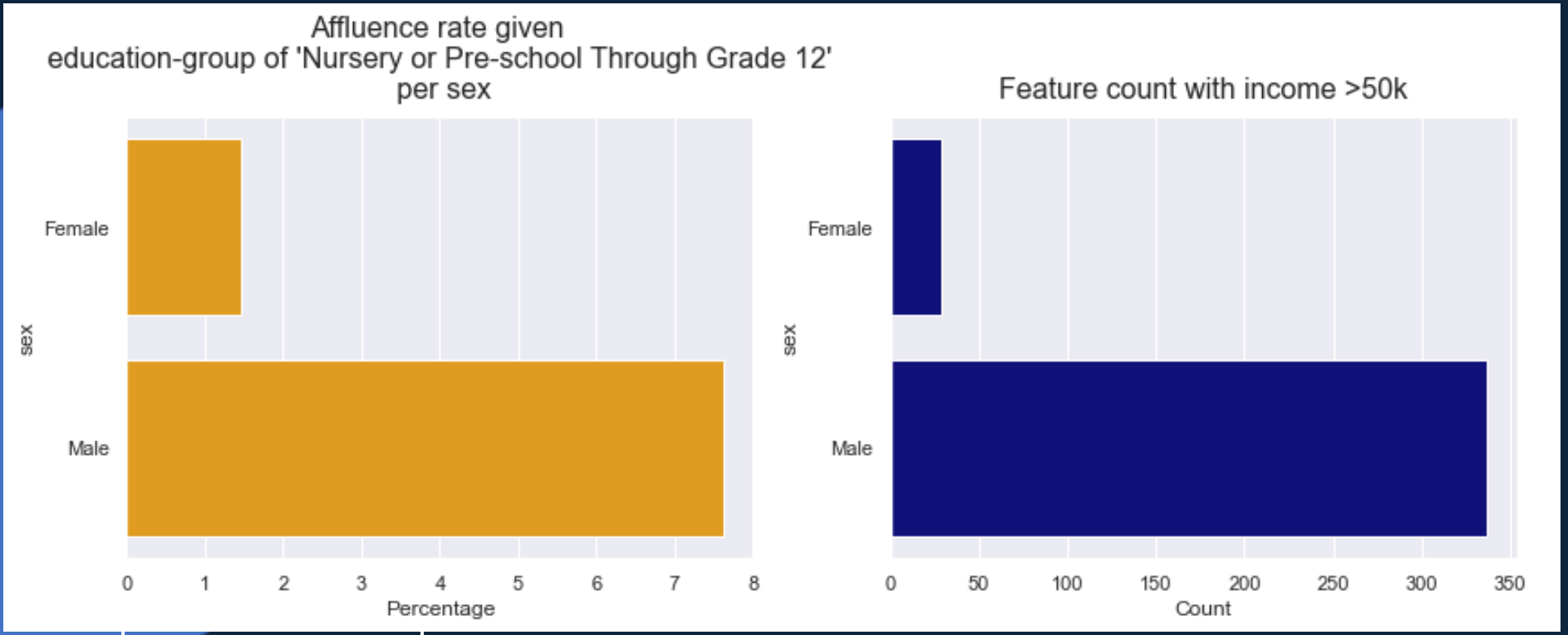
Bivariate Analysis: Education vs Sex

Males are more likely to have an income >50k than *females, g*iven the same educational attainment.



Bivariate Analysis: Education vs Sex

Males are more likely to have an income >50k than *females, g*iven the same educational attainment.



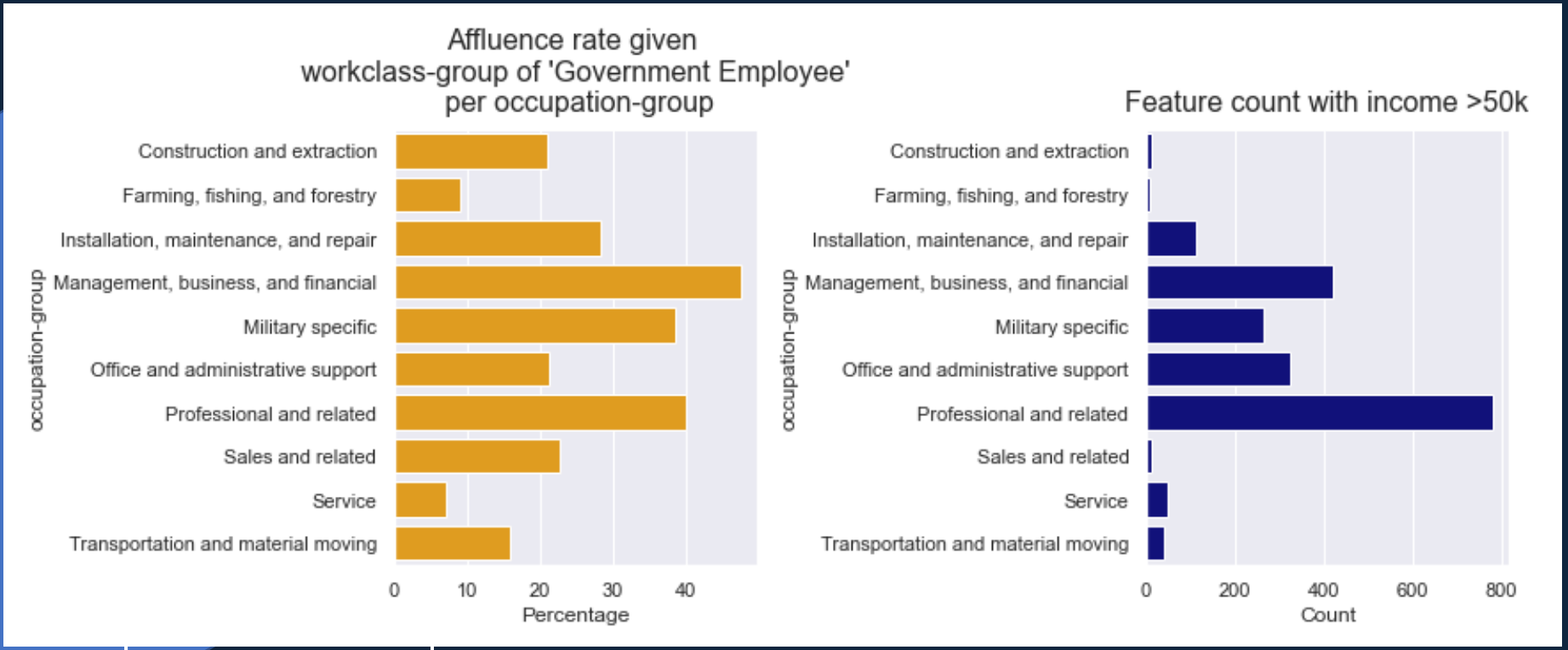
Work class vs Occupation

- Government Employee
- Private Sector Employee
- Self-employed or Other
- Unknown

- Construction and Extraction
- Farming, fishing and forestry
- Installation, maintenance, and repair
- Management, business, and financial
- Military Specific
- Office and administrative support
- Professional and related
- Sales and related
- Service
- Transportation and material moving
- Unknown

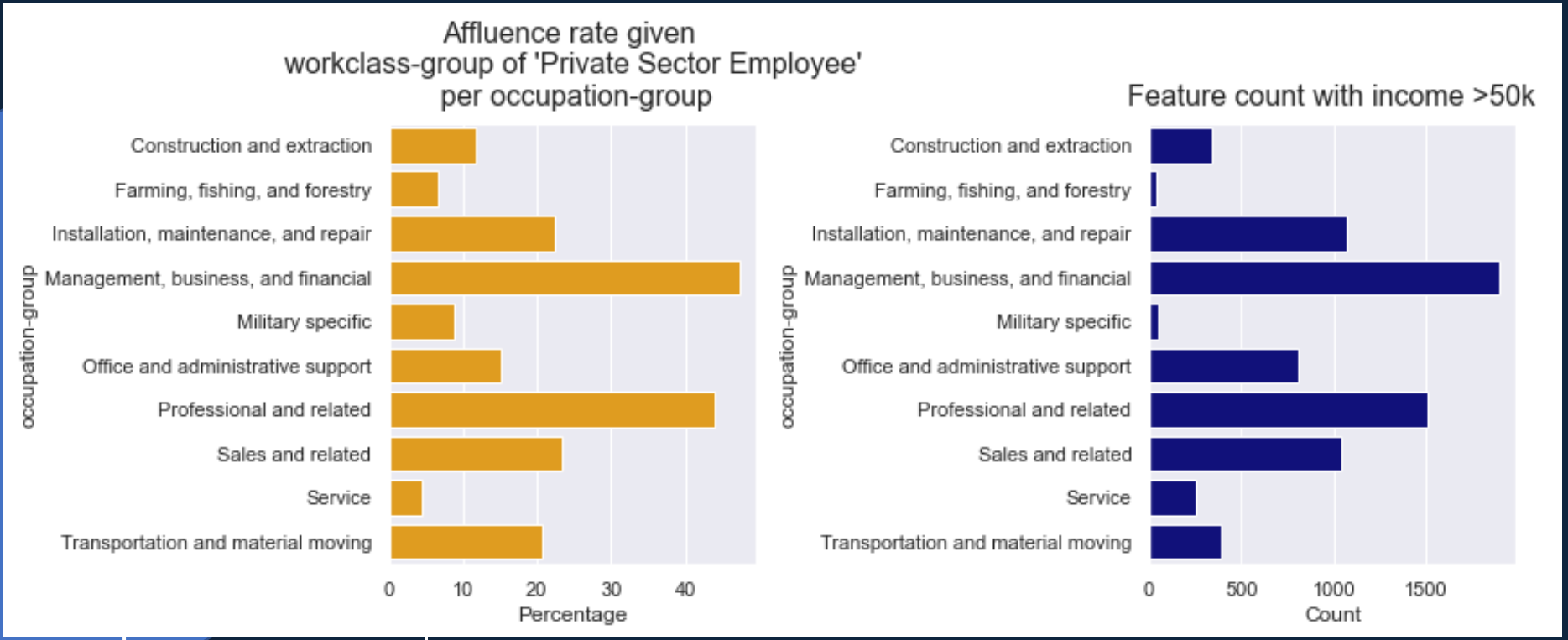
Bivariate Analysis: Work class vs Occupation

The highest-paid jobs in the government are in *management, professionals,* and being in the *military.*



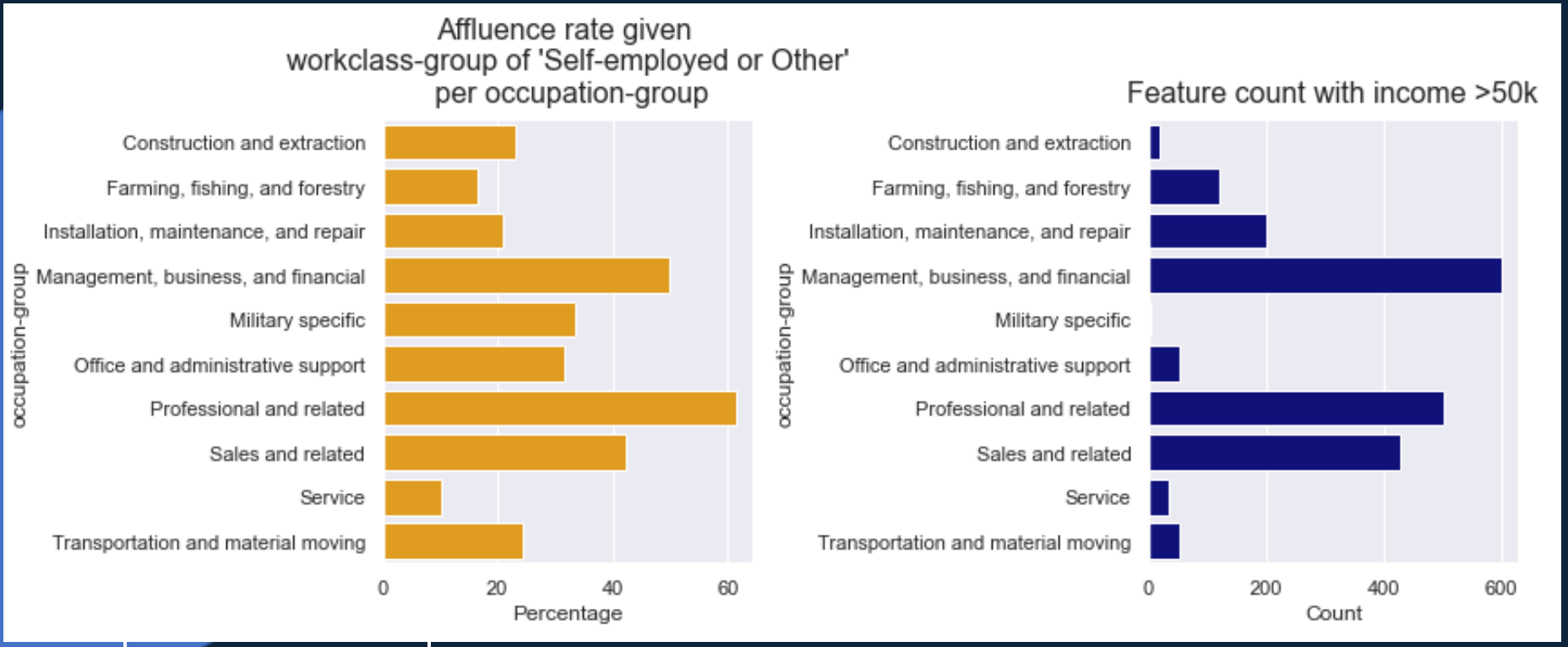
Bivariate Analysis: Work class vs Occupation

Management, professionals, and *sales* are still the top occupations for non-government employees.



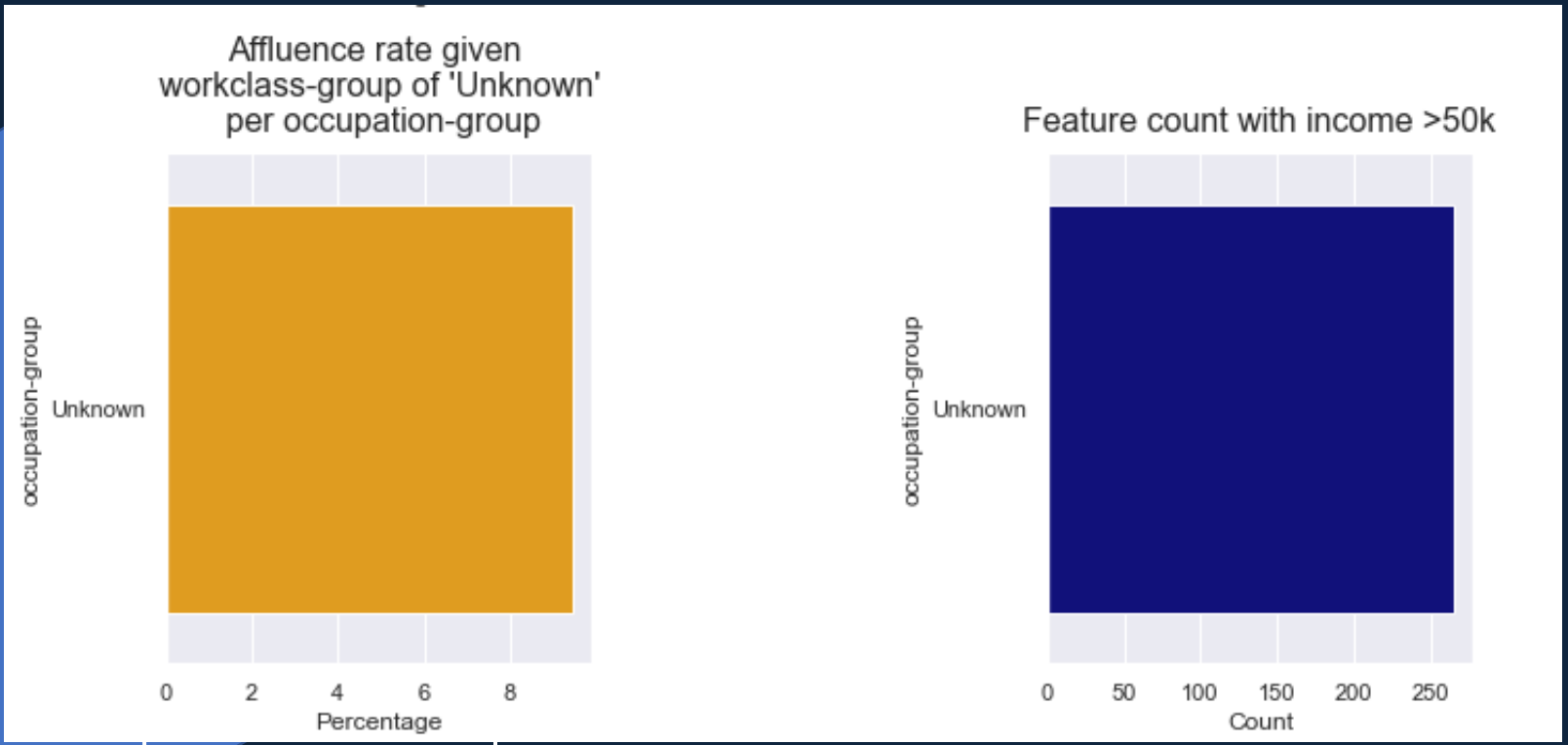
Bivariate Analysis: Work class vs Occupation

Management, professionals, and *sales* are still the top occupations for non-government employees.



Bivariate Analysis: Work class vs Occupation

Surprisingly, there are *unknown* jobs under *unknown* class groups that are earning >50k.



Work Class vs Work Hours

- After Bachelor's Degree
- College or Some College
- High School
- Nursery or Pre-school through Grade 12

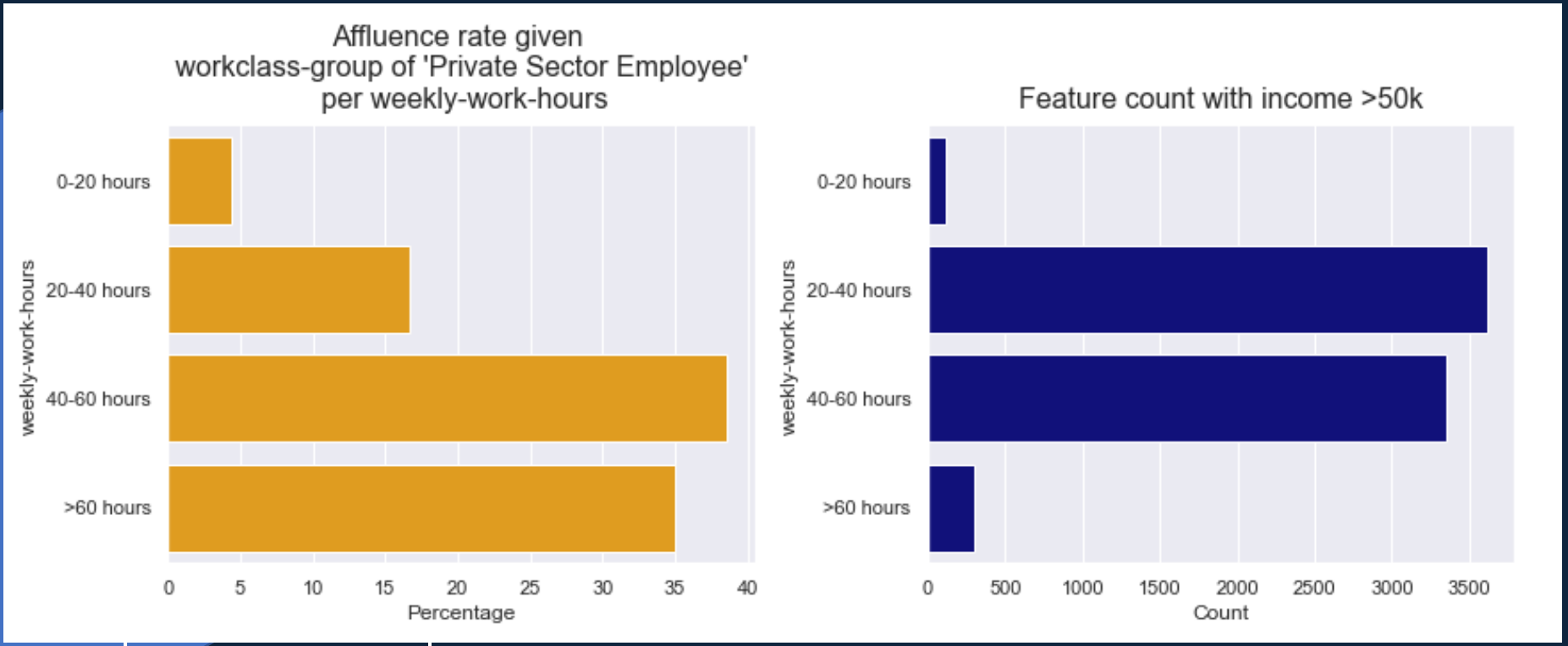
- 0 - 20 hours
- 20 – 40 hours
- 40 to 60 hours
- >60 hours

Bivariate Analysis: Work Class vs Work Hours

In comparing *government* and *self-employment*, working *20-40 hours* benefits the self-employed, but the dynamic reverses for *40-60 hours*.

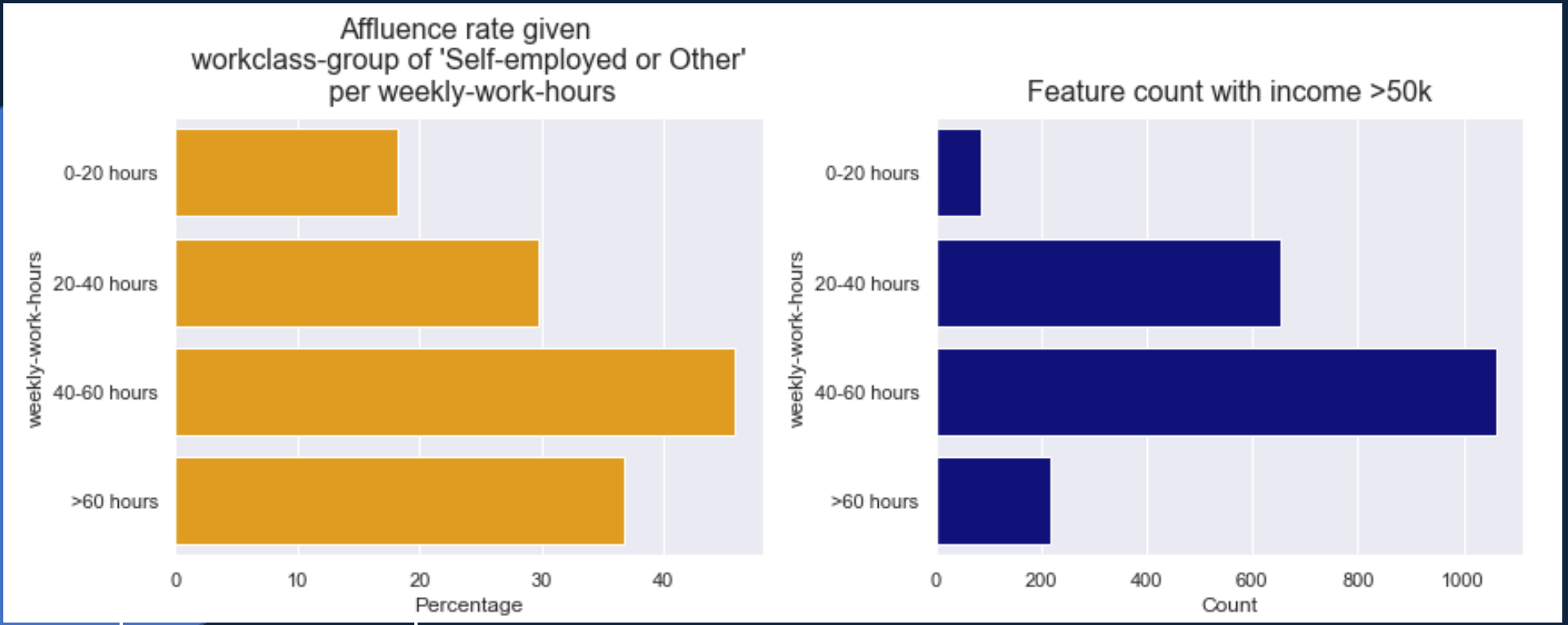
Bivariate Analysis: Work Class vs Work Hours

In comparing *government* and *self-employment*, working *20-40 hours* benefits the self-employed, but the dynamic reverses for *40-60 hours*.



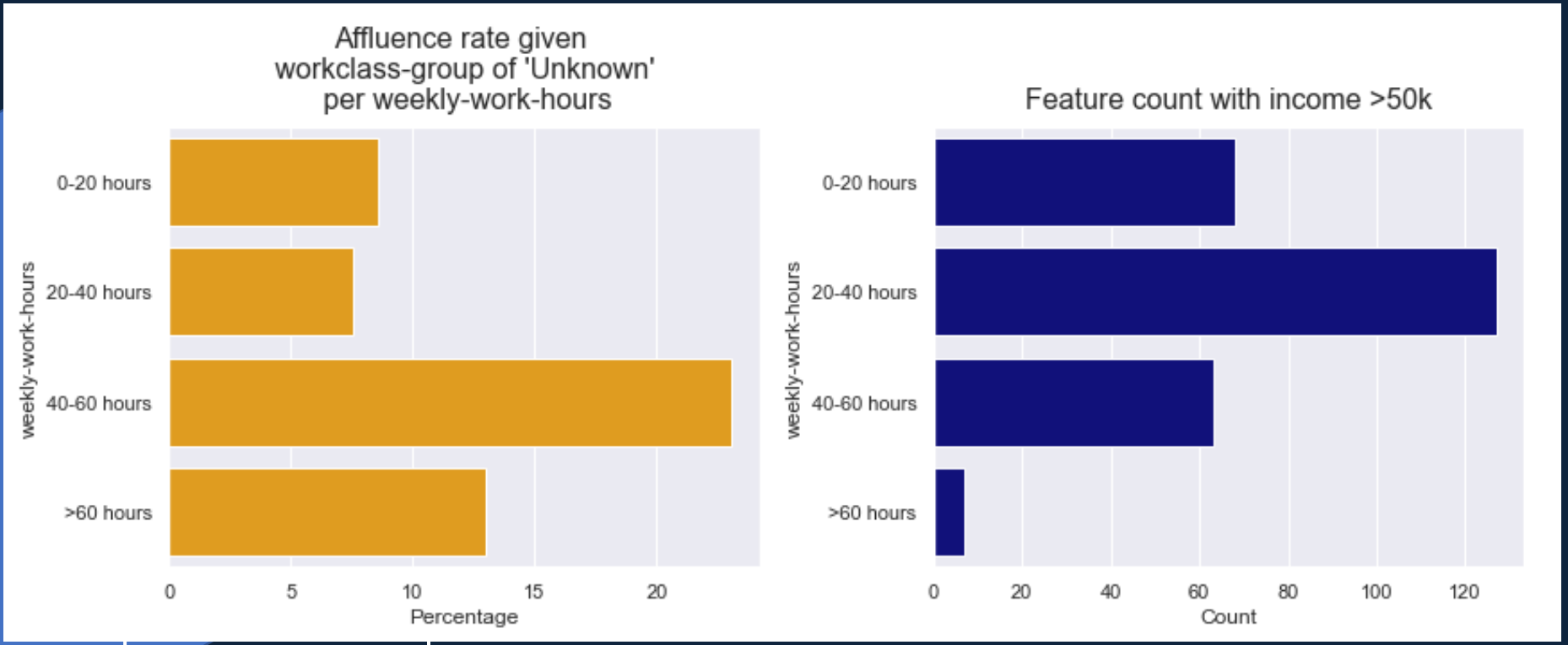
Bivariate Analysis: Work Class vs Work Hours

Private sector employees should work more than *40 hours* to have an increased probability of having an income >50k.



Bivariate Analysis: Work Class vs Work Hours

Working *0-20 hours is* much more beneficial than working *20-40 hours* in *unknown* work class groups.



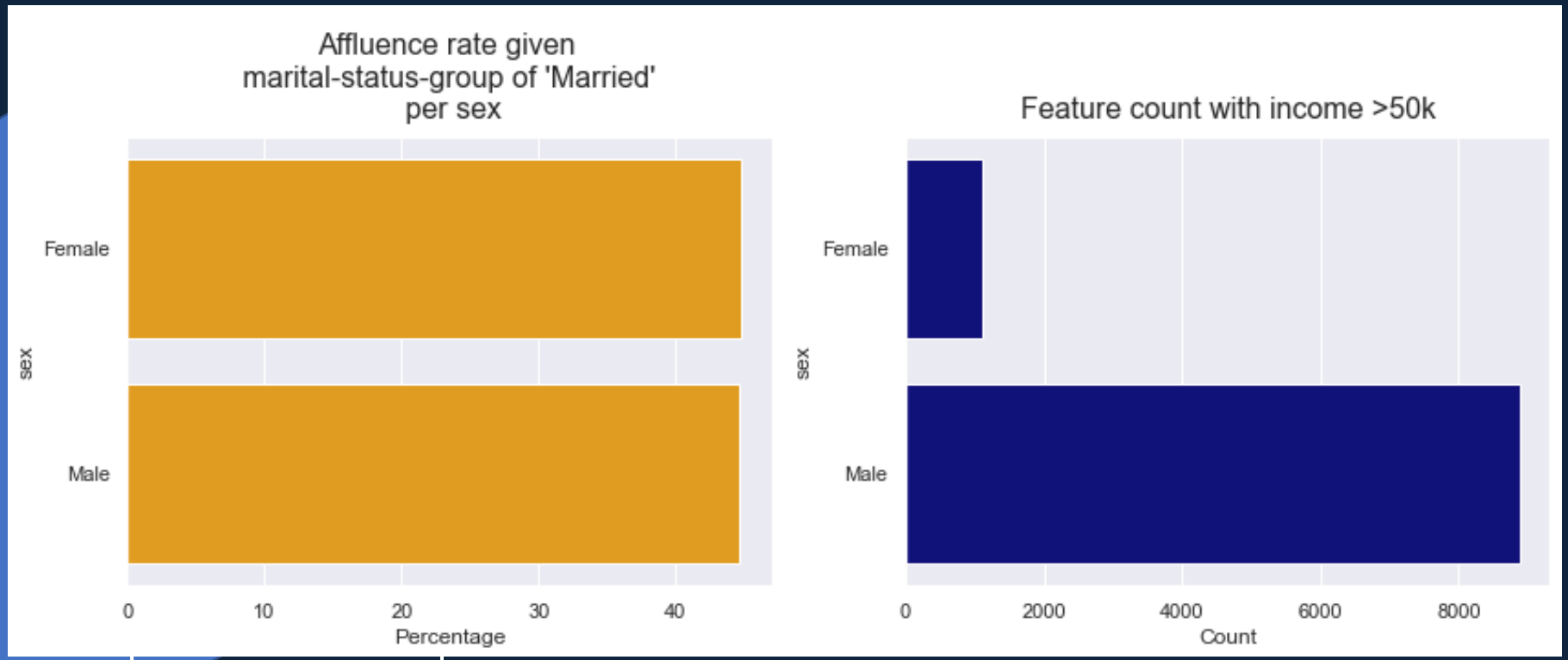
Marital Status vs Sex

- Married
- Never-married
- Separated
- Widowed

- Female
- Male

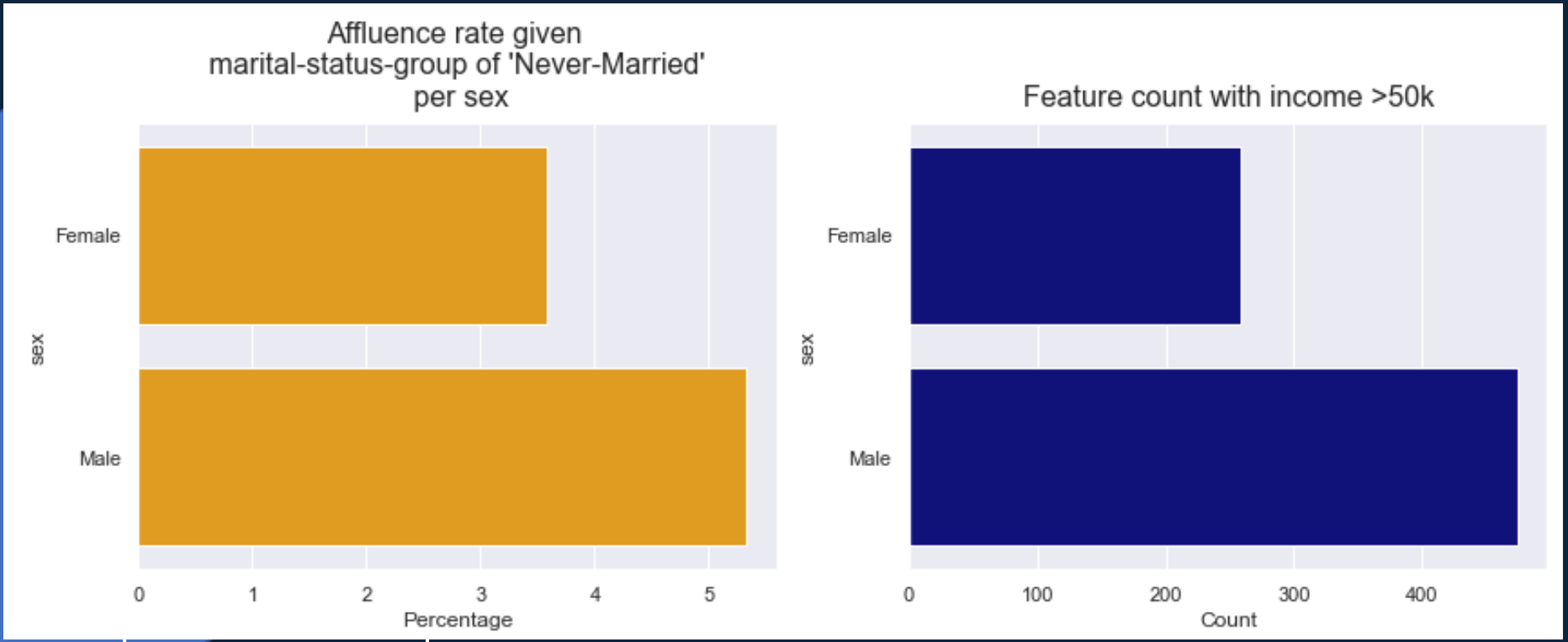
Bivariate Analysis: Marital Status vs Sex

For *married* and *never-married* people, *male* and *female have equal chances.*



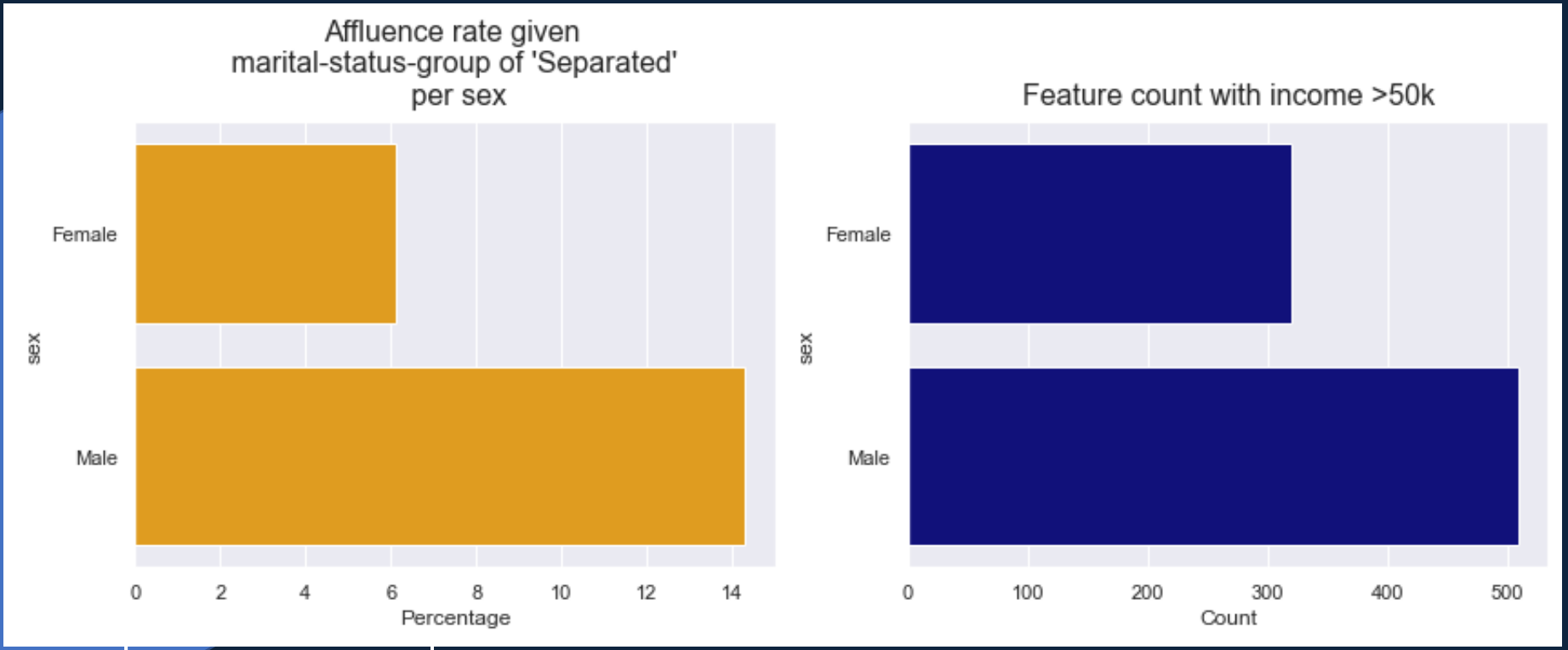
Bivariate Analysis: Marital Status vs Sex

For *married* and *never-married* people, *male* and *female have equal chances.*



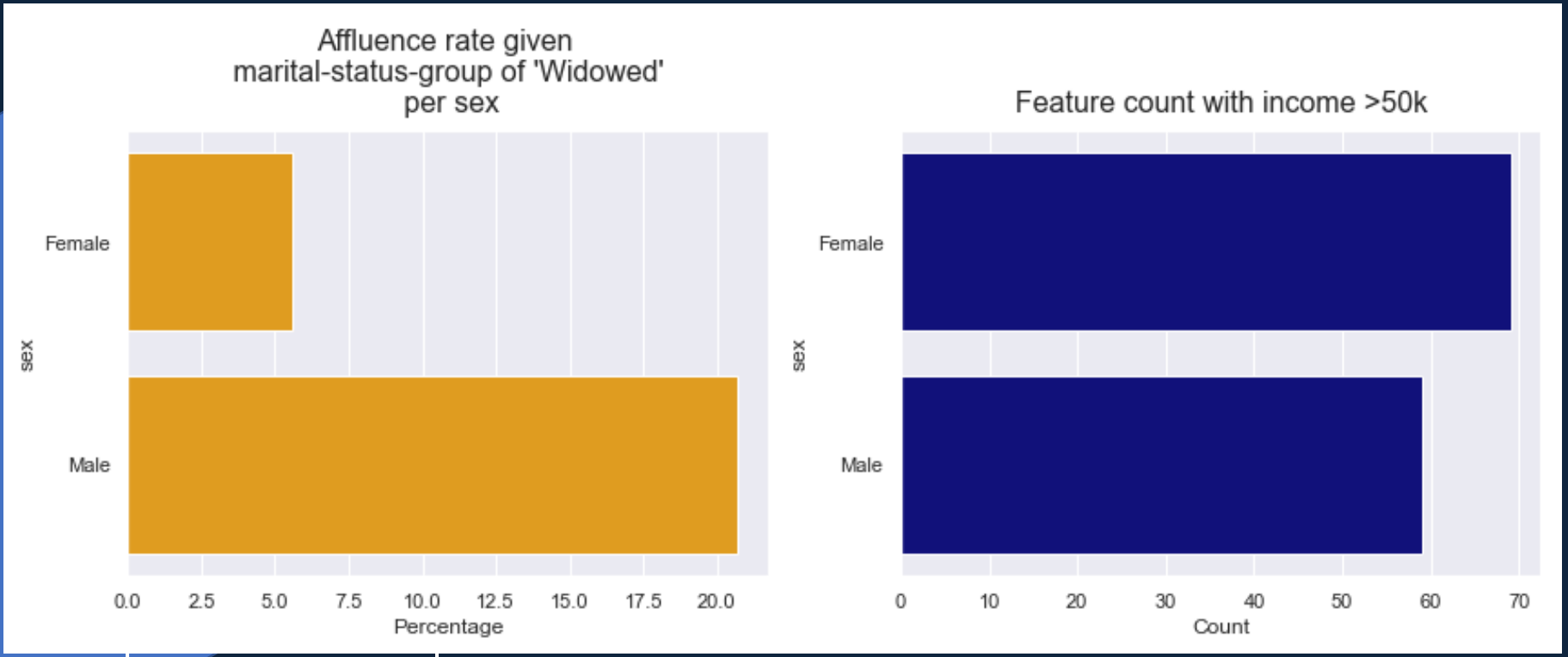
Bivariate Analysis: Marital Status vs Sex

For *separated* and *widowed* people, *males* have a greater probability of having an income >50k.



Bivariate Analysis: Marital Status vs Sex

For *separated* and *widowed* people, *males* have a greater probability of having an income >50k.



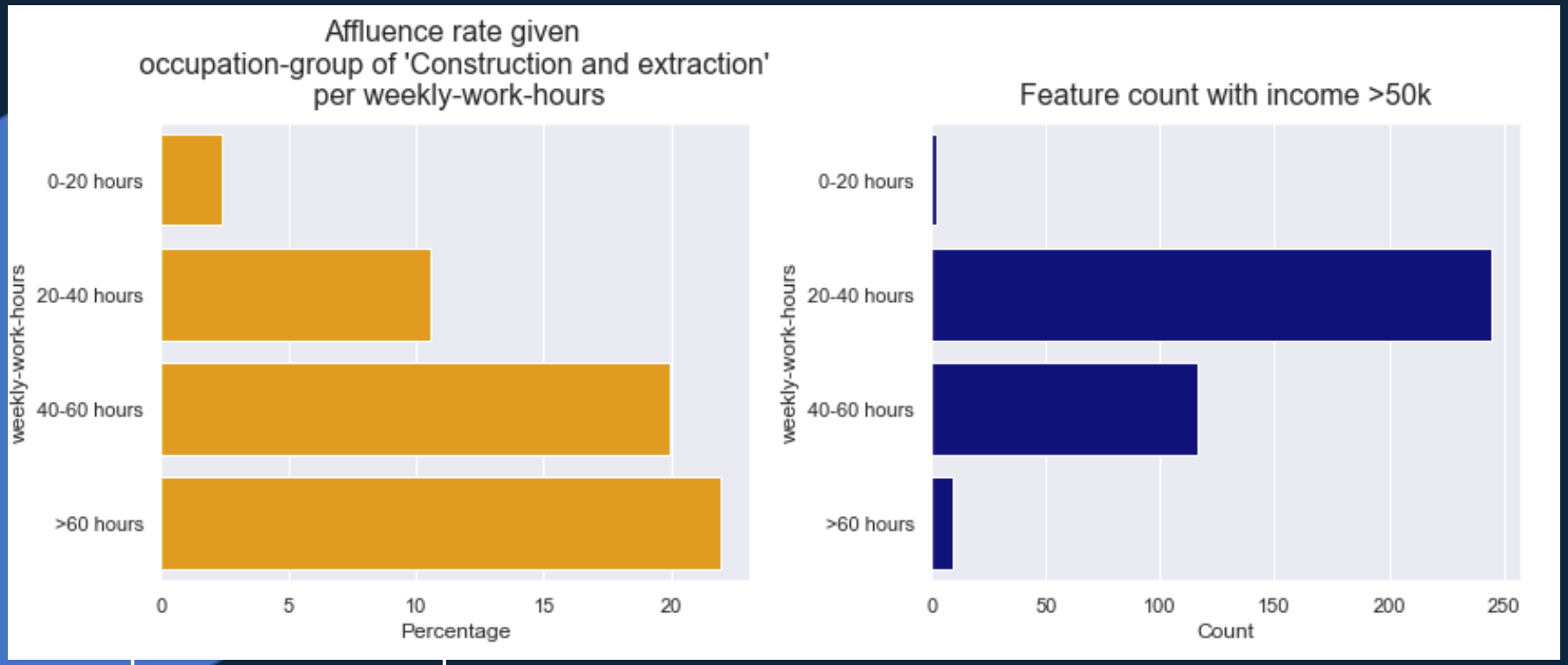
Occupation vs Work Hours

- Construction and Extraction
- Farming, fishing and forestry
- Installation, maintenance, and repair
- Management, business, and financial
- Military Specific
- Office and administrative support
- Professional and related
- Sales and related
- Service
- Transportation and material moving
- Unknown

- 0 - 20 hours
- 20 – 40 hours
- 40 to 60 hours
- >60 hours

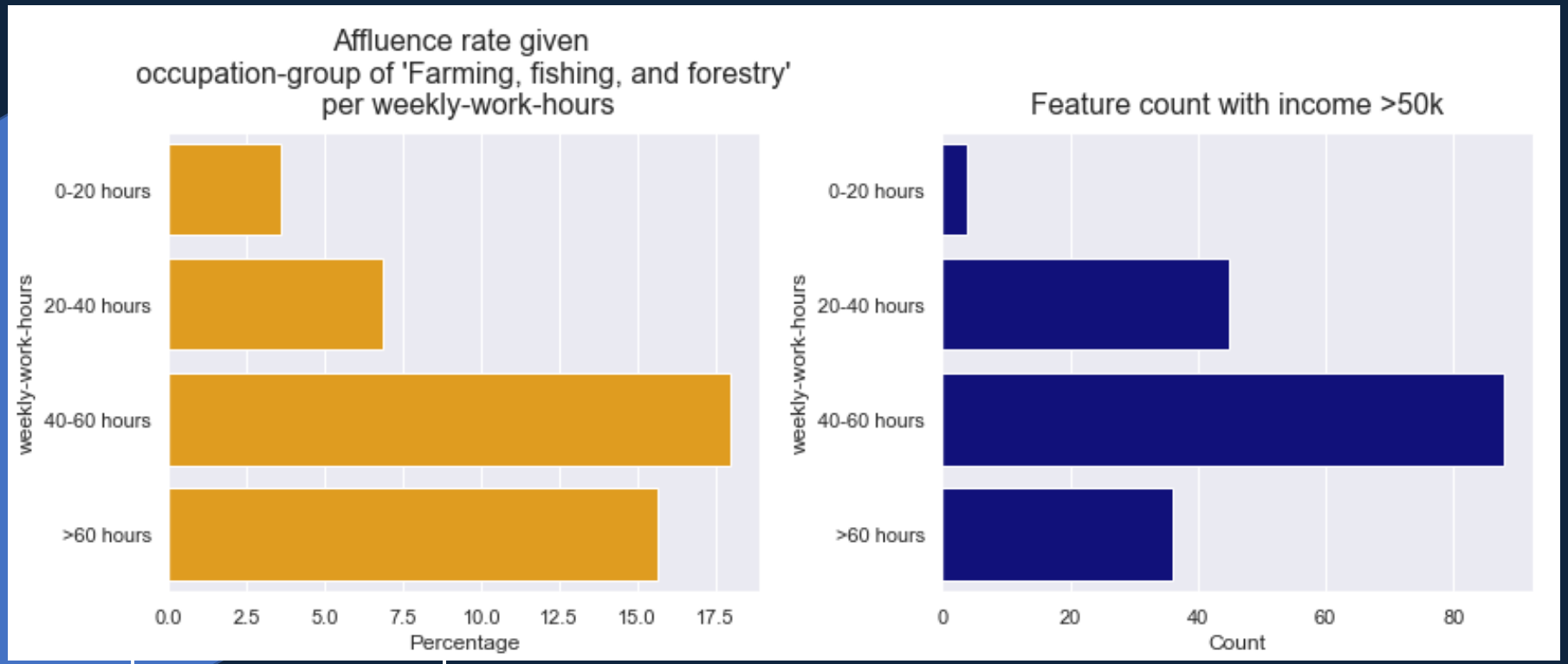
Bivariate Analysis: Occupation vs Work hours

Construction, military, office and adm. support and *transportation and material moving* are the only jobs that have a higher chance for >60 hours.



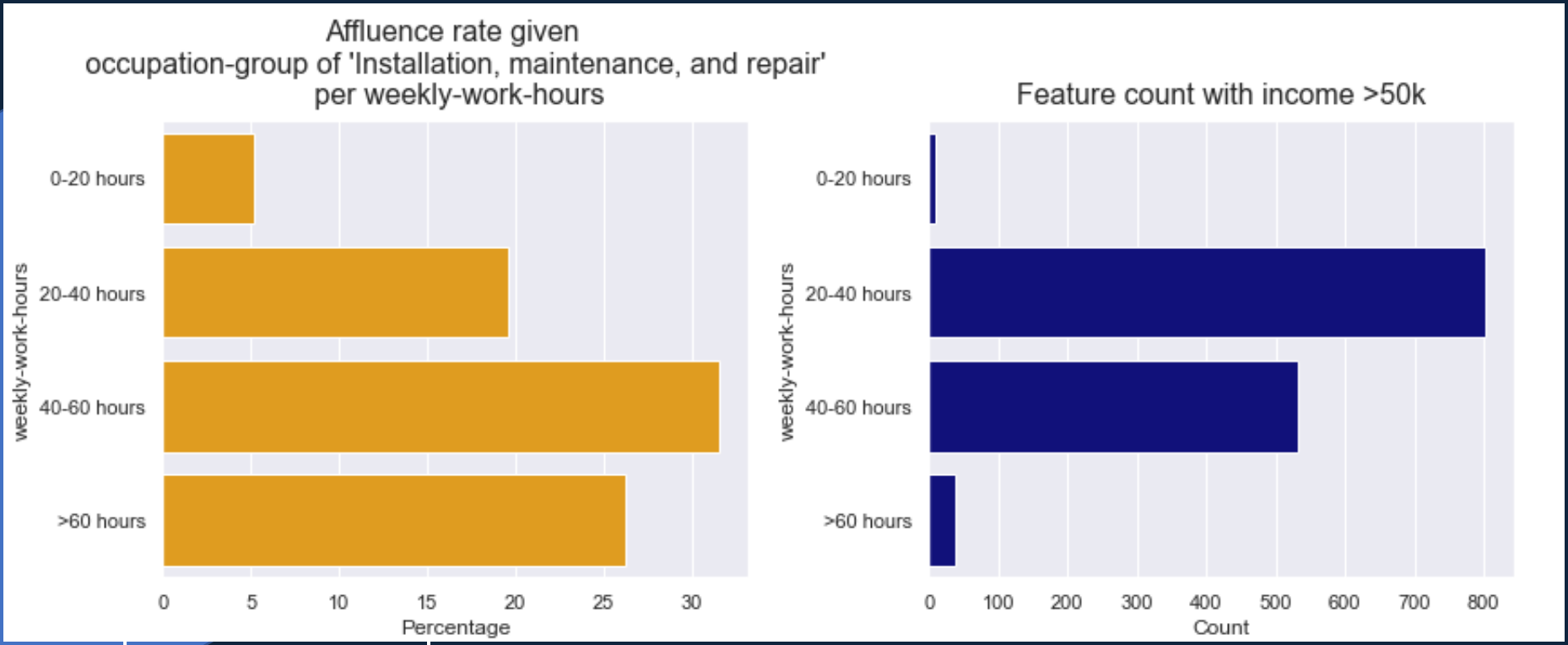
Bivariate Analysis: Occupation vs Work hours

Construction, military, office and adm. support and *transportation and material moving* are the only jobs that have a higher chance for >60 hours.



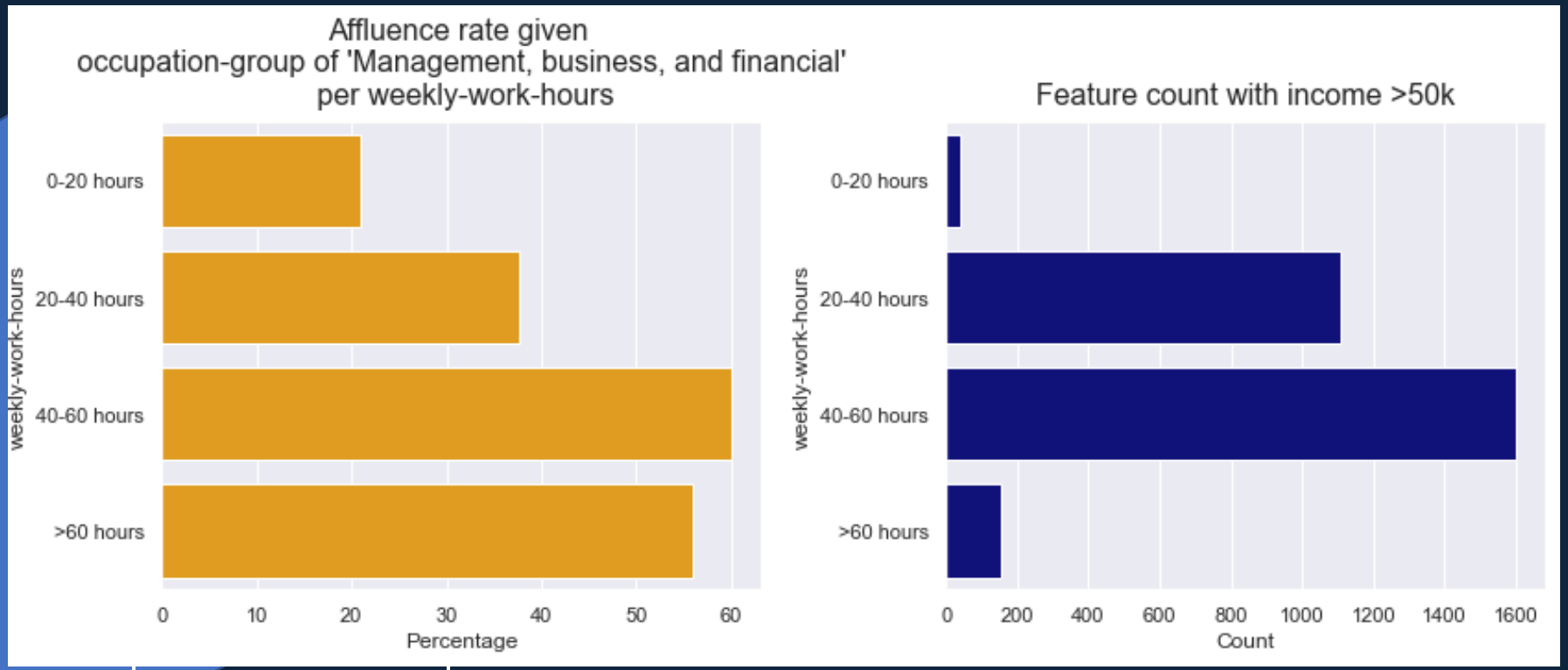
Bivariate Analysis: Occupation vs Work hours

Construction, military, office and adm. support and *transportation and material moving* are the only jobs that have a higher chance for >60 hours.



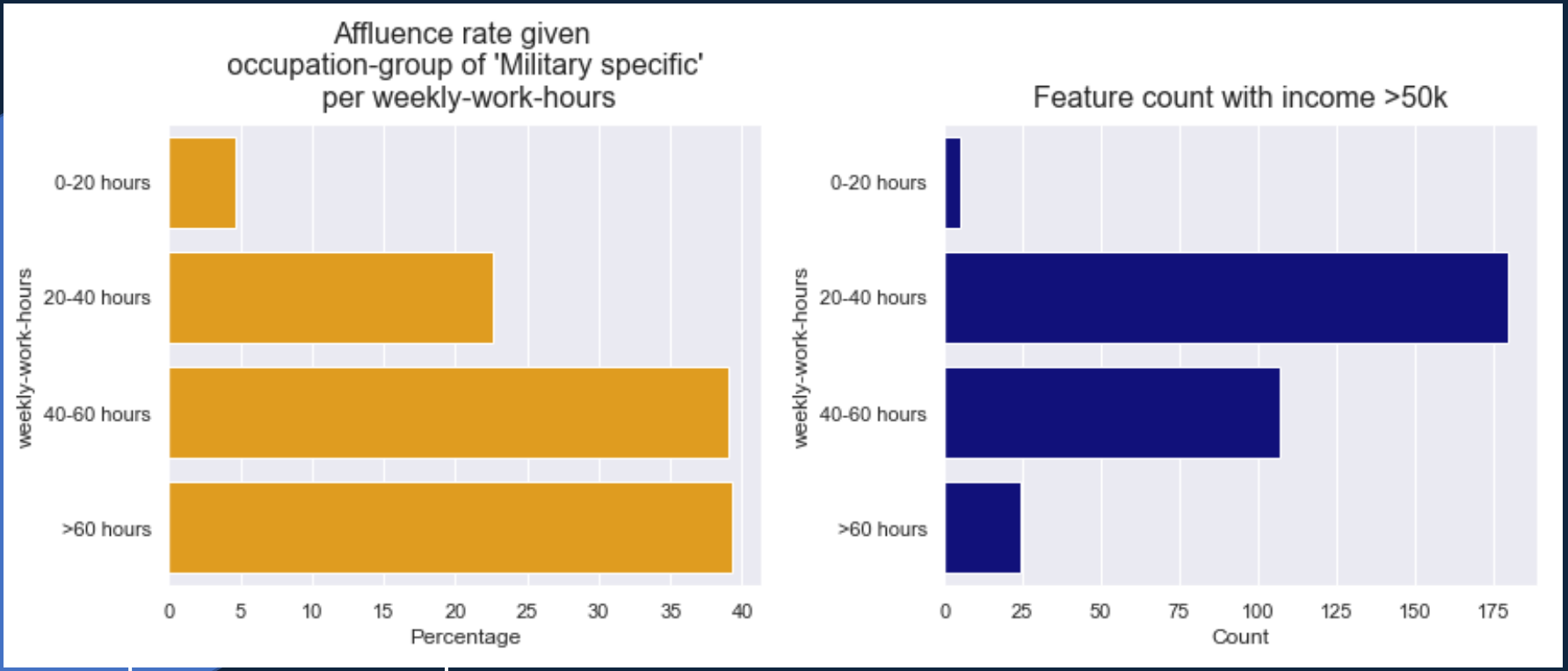
Bivariate Analysis: Occupation vs Work hours

Construction, military, office and adm. support and *transportation and material moving* are the only jobs that have a higher chance for >60 hours.



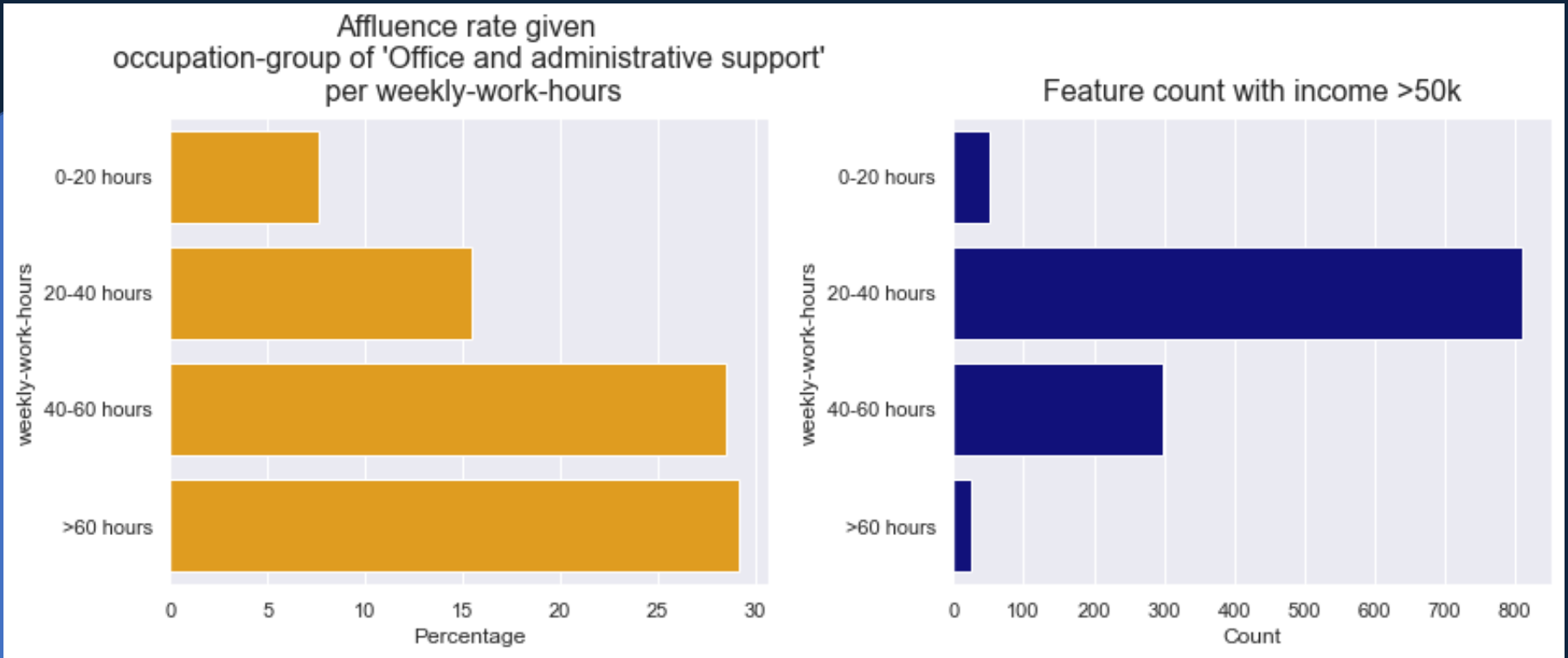
Bivariate Analysis: Occupation vs Work hours

Construction, military, office and adm. support and *transportation and material moving* are the only jobs that have a higher chance for >60 hours.



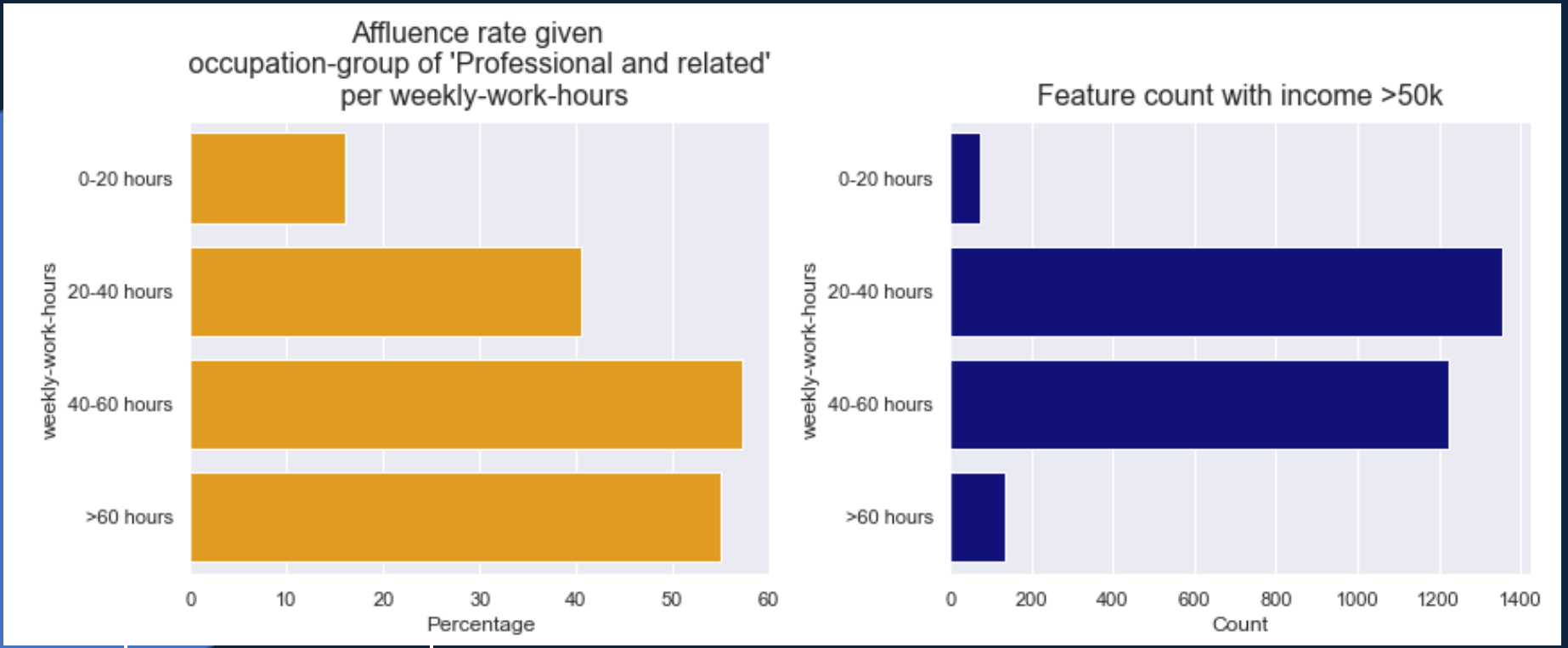
Bivariate Analysis: Occupation vs Work hours

Construction, military, office and adm. support and *transportation and material moving* are the only jobs that have a higher chance for >60 hours.



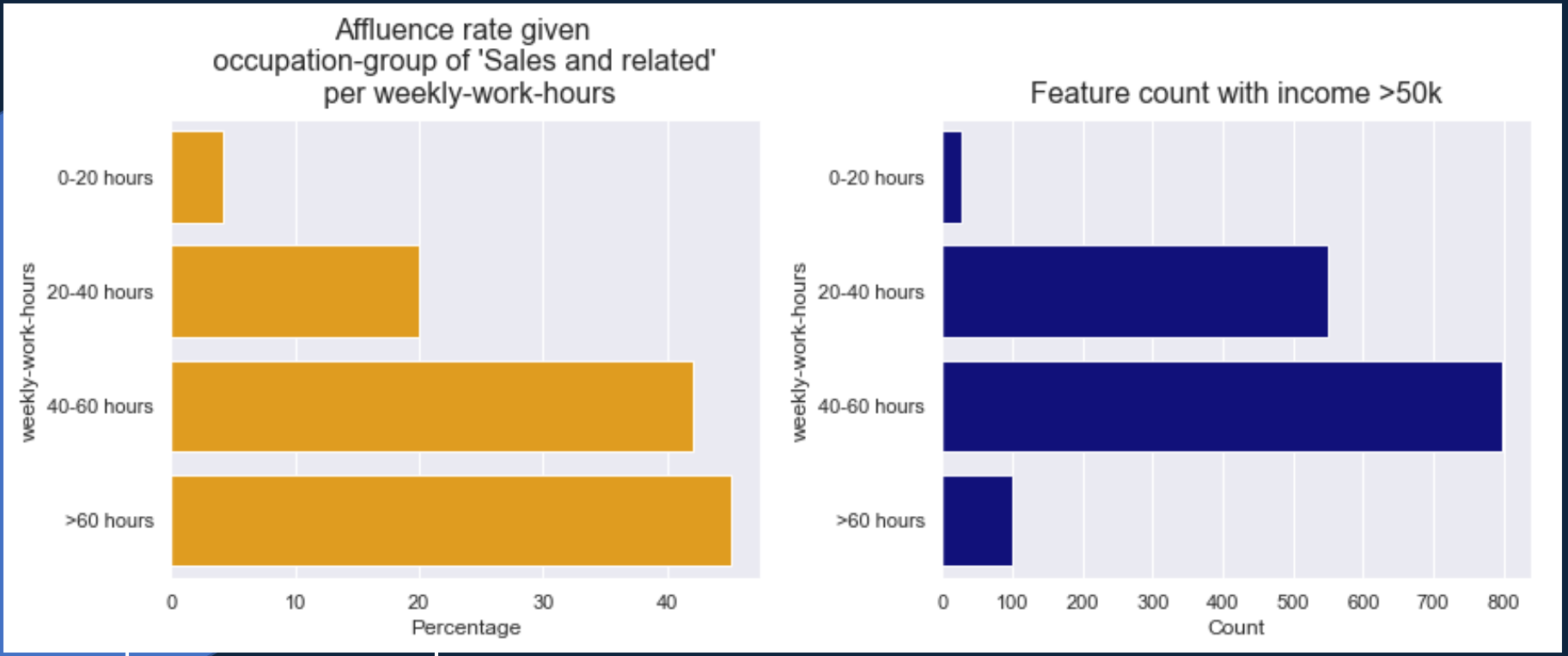
Bivariate Analysis: Occupation vs Work hours

Construction, military, office and adm. support and *transportation and material moving* are the only jobs that have a higher chance for >60 hours.



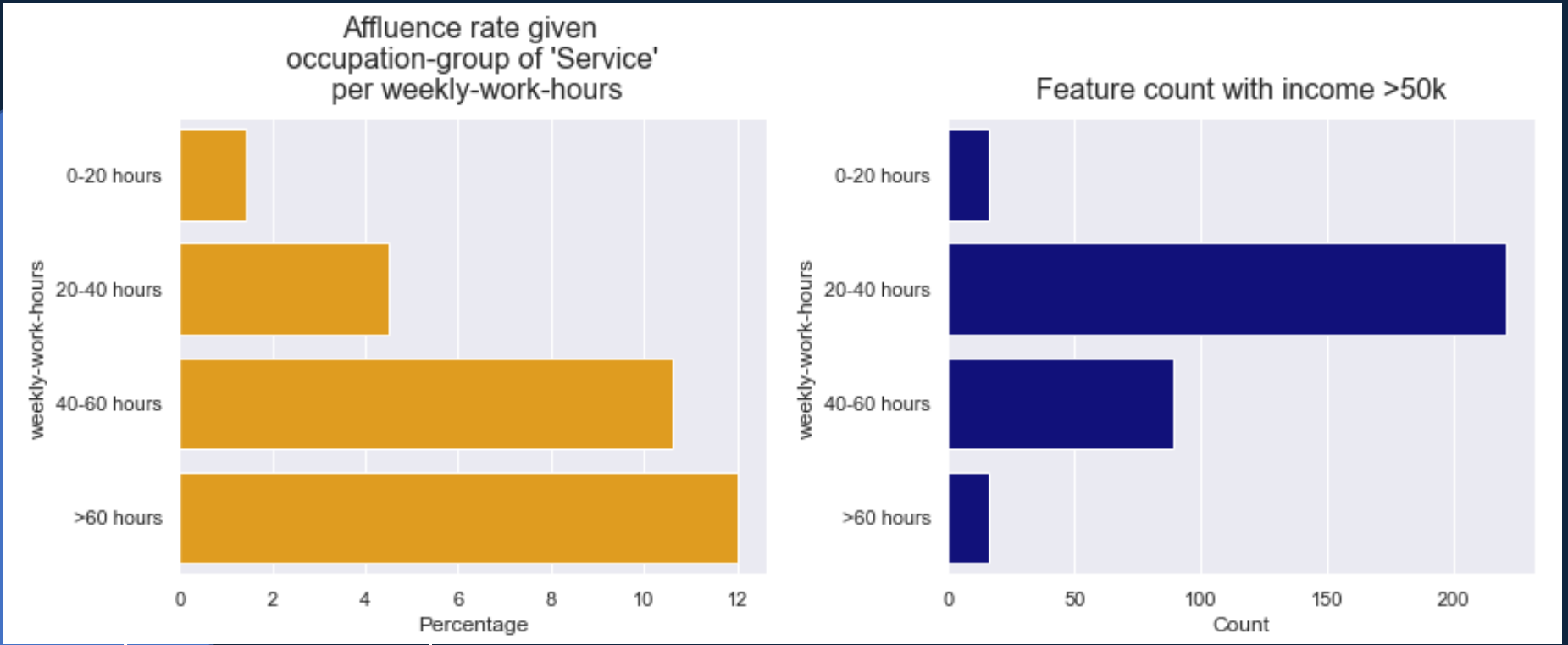
Bivariate Analysis: Occupation vs Work hours

Construction, military, office and adm. support and *transportation and material moving* are the only jobs that have a higher chance for >60 hours.



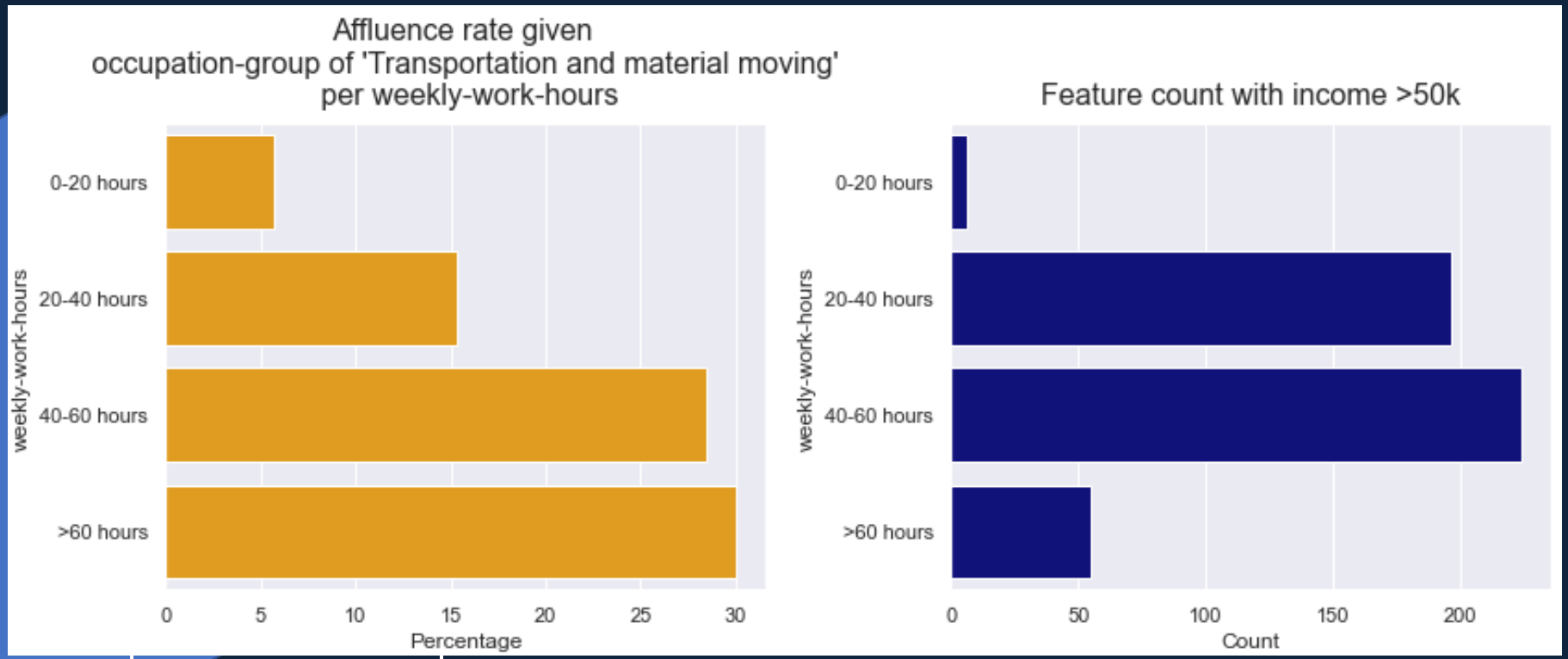
Bivariate Analysis: Occupation vs Work hours

Construction, military, office and adm. support and *transportation and material moving* are the only jobs that have a higher chance for >60 hours.



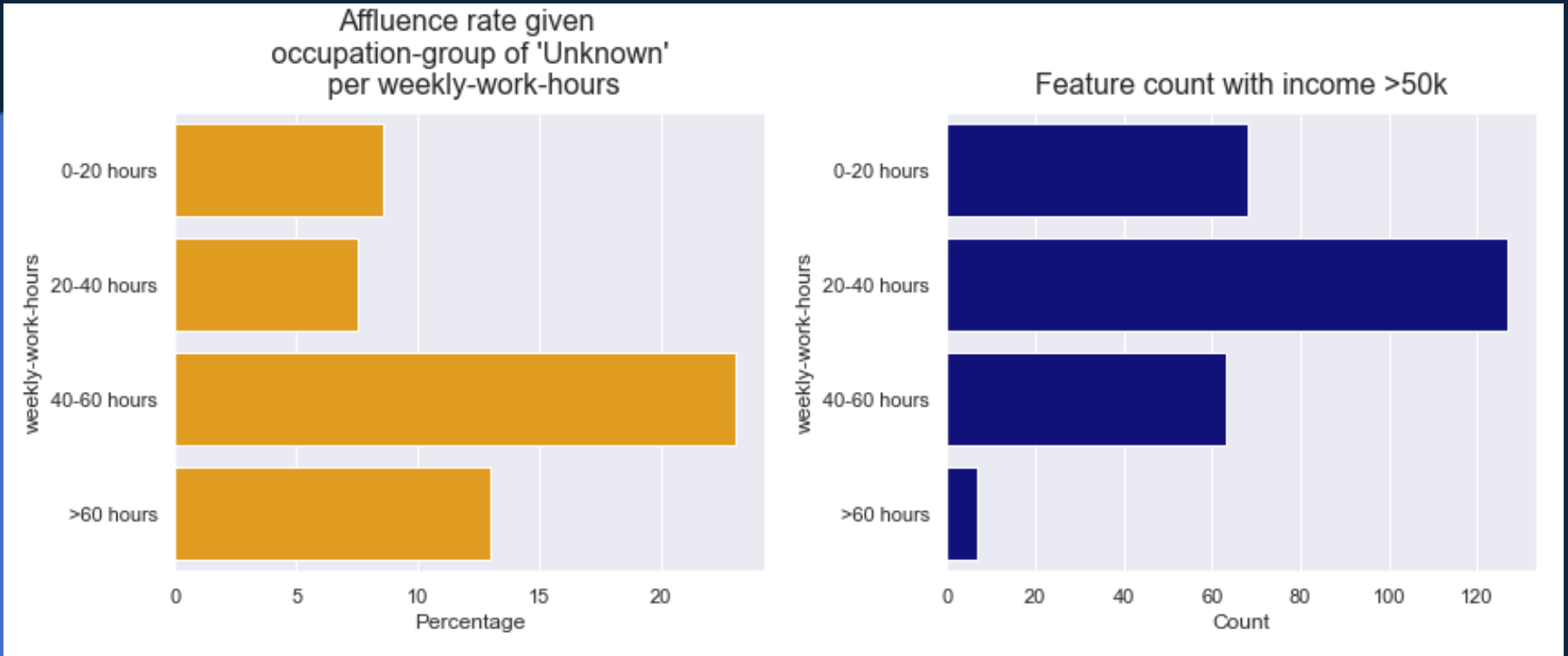
Bivariate Analysis: Occupation vs Work hours

Construction, military, office and adm. support and *transportation and material moving* are the only jobs that have a higher chance for >60 hours.



Bivariate Analysis: Occupation vs Work hours

Construction, military, office and adm. support and *transportation and material moving* are the only jobs that have a higher chance for >60 hours.



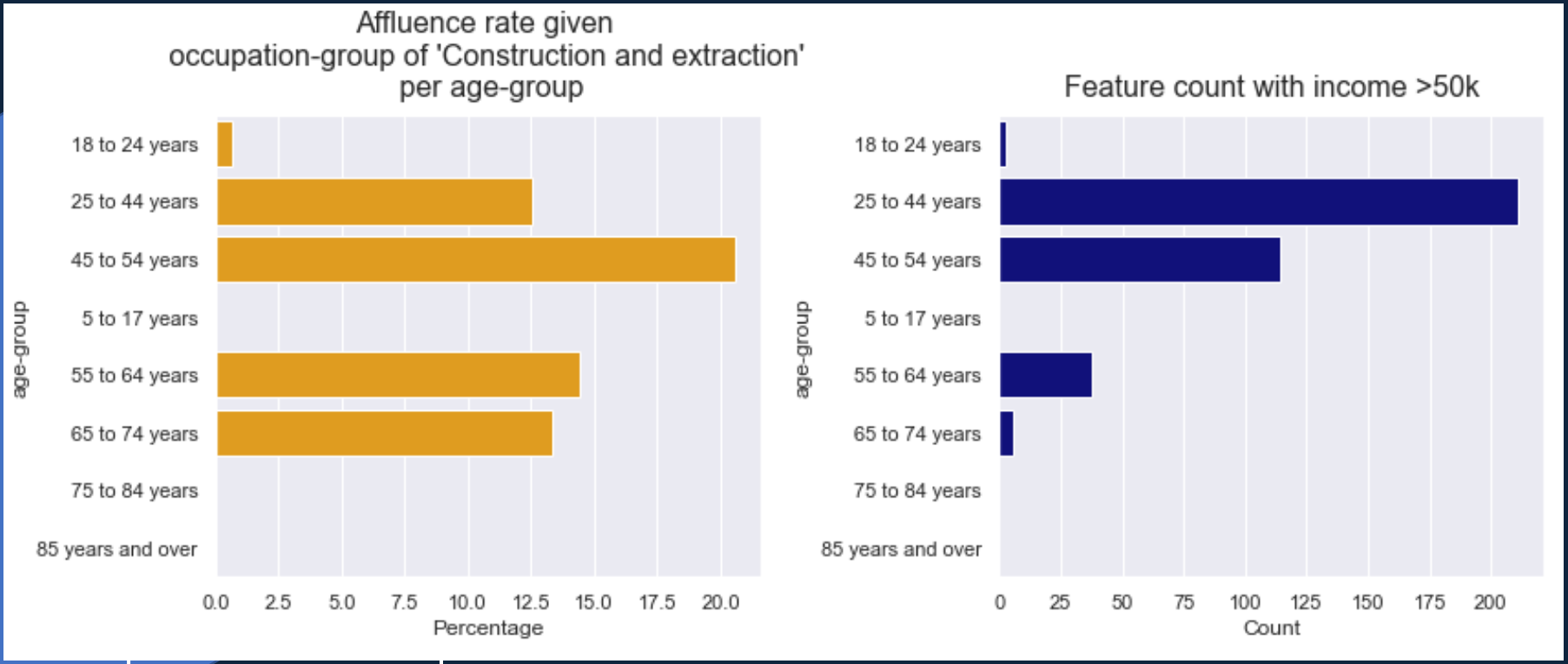
Occupation vs Age

- Construction and Extraction
- Farming, fishing and forestry
- Installation, maintenance, and repair
- Management, business, and financial
- Military Specific
- Office and administrative support
- Professional and related
- Sales and related
- Service
- Transportation and material moving
- Unknown

- 5 to 17 years
- 18 to 24 years
- 25 to 44 years
- 44 to 54 years
- 55 to 64 years
- 65 to 74 years
- 75 to 84 years
- 85 years and over

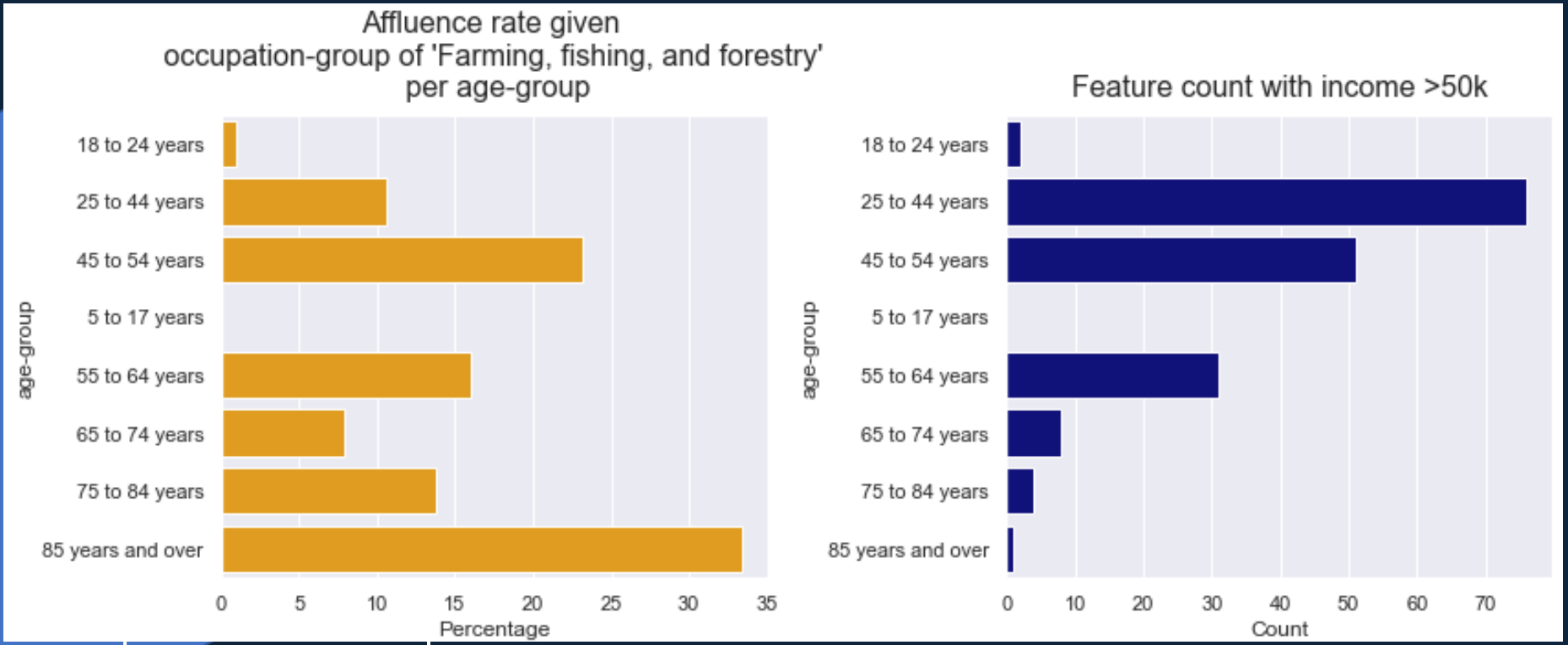
Bivariate Analysis: Occupation vs Age

Regardless of occupation, the probability peaks in the *45 to 54 years* old age group.



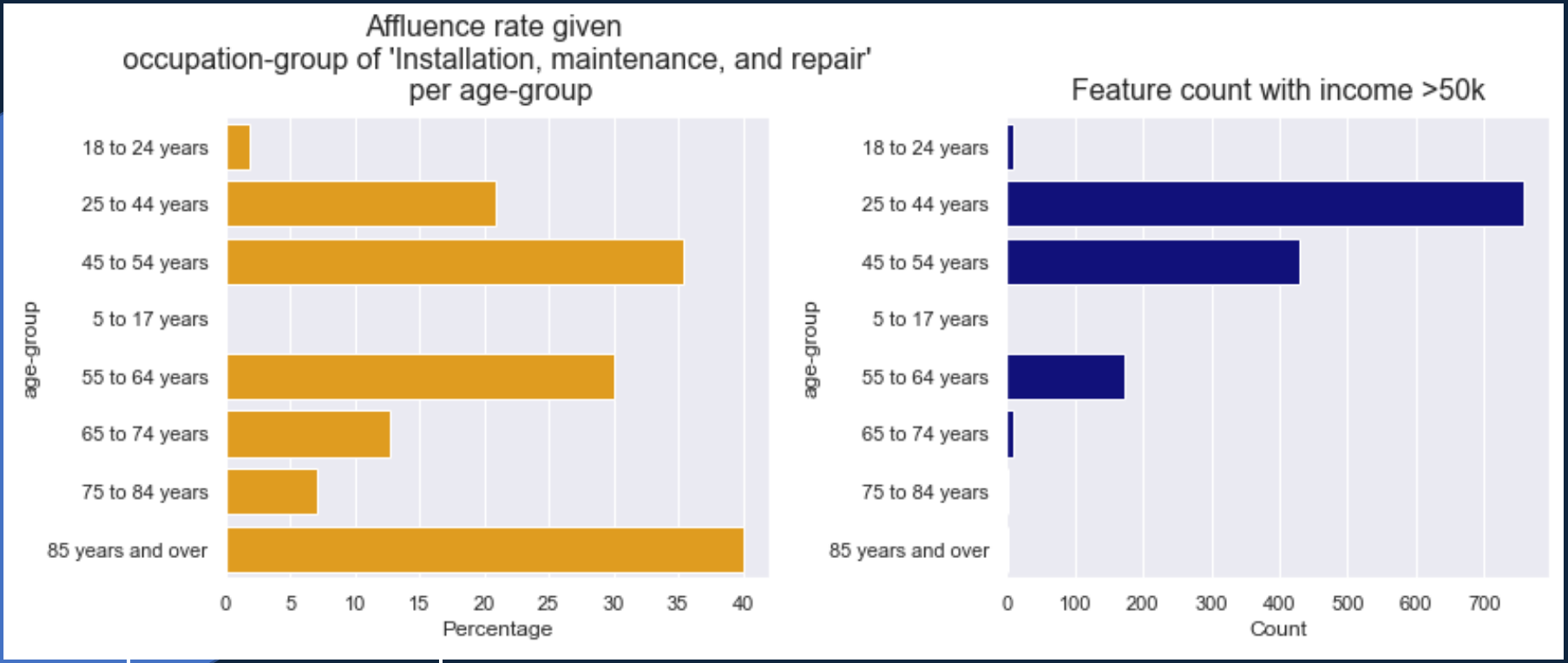
Bivariate Analysis: Occupation vs Age

Regardless of occupation, the probability peaks in the *45 to 54 years* old age group.



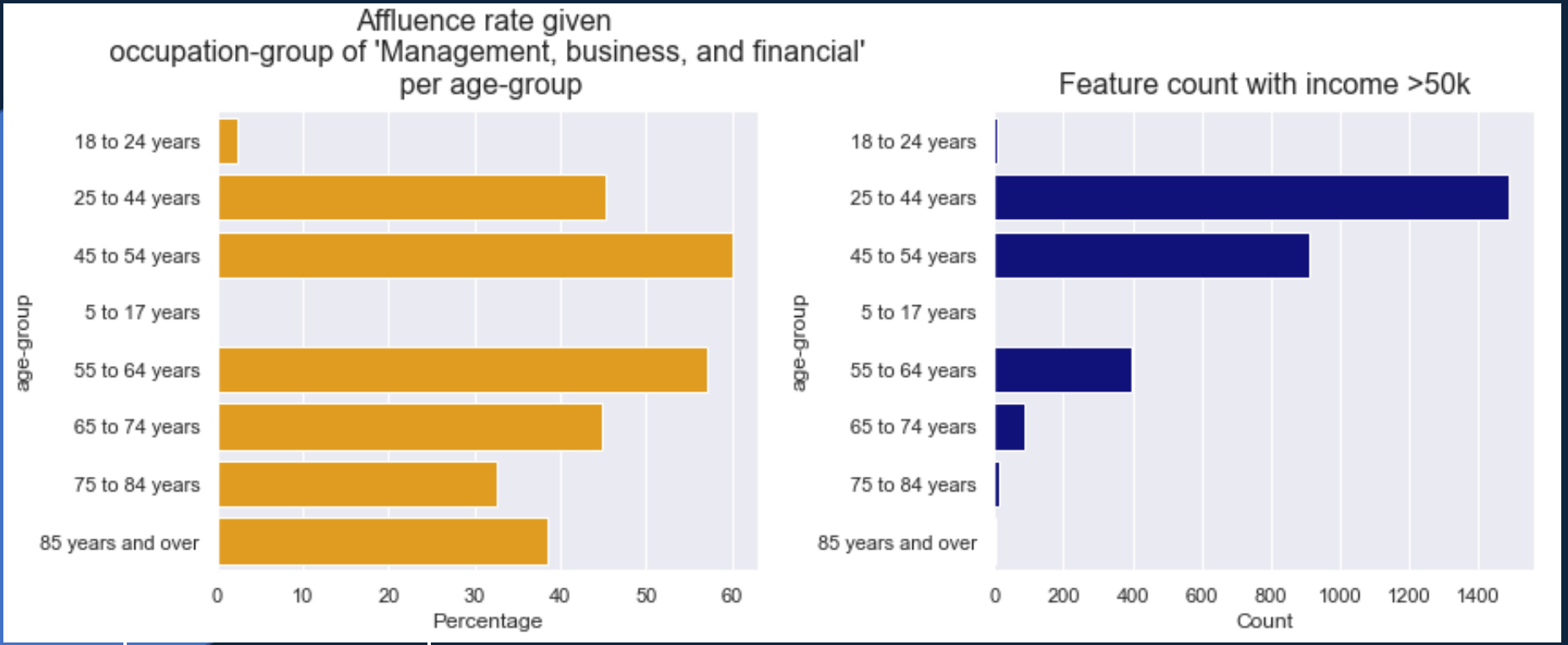
Bivariate Analysis: Occupation vs Age

Regardless of occupation, the probability peaks in the *45 to 54 years* old age group.



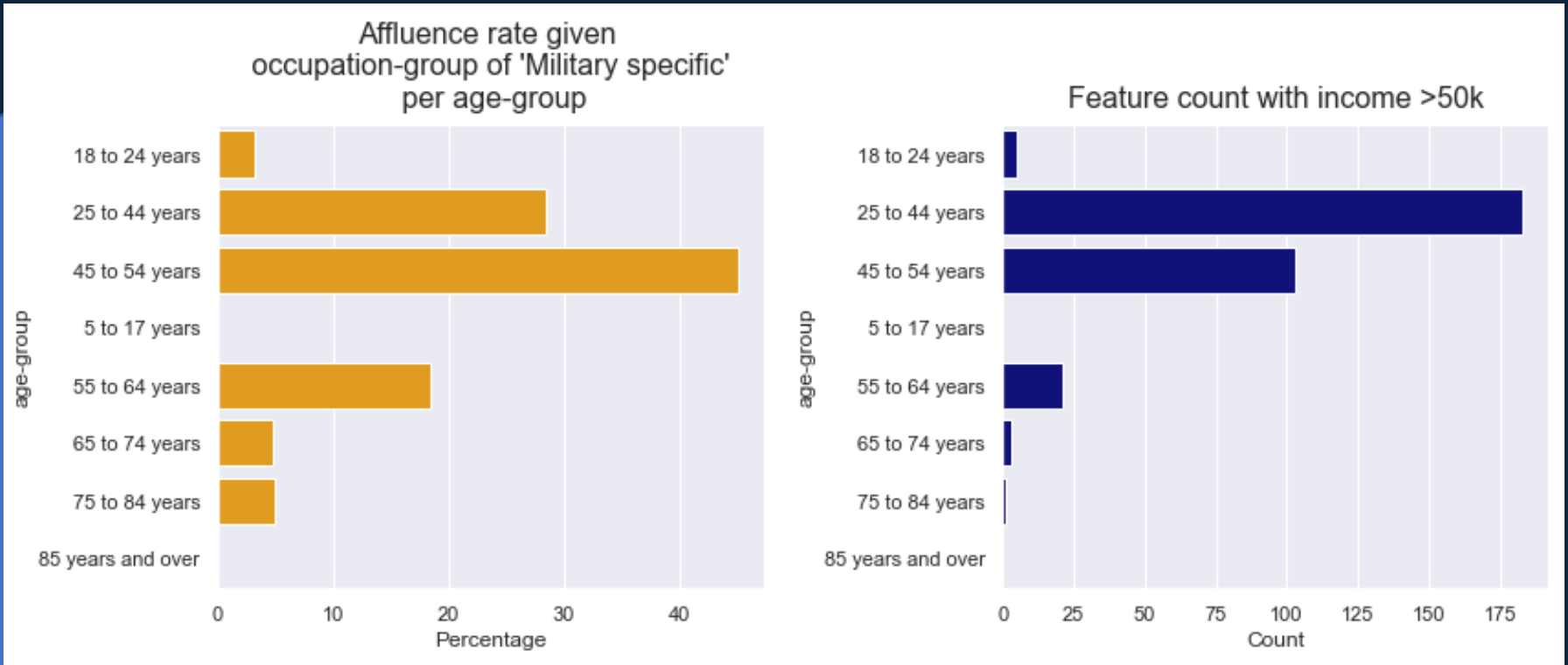
Bivariate Analysis: Occupation vs Age

Regardless of occupation, the probability peaks in the *45 to 54 years* old age group.



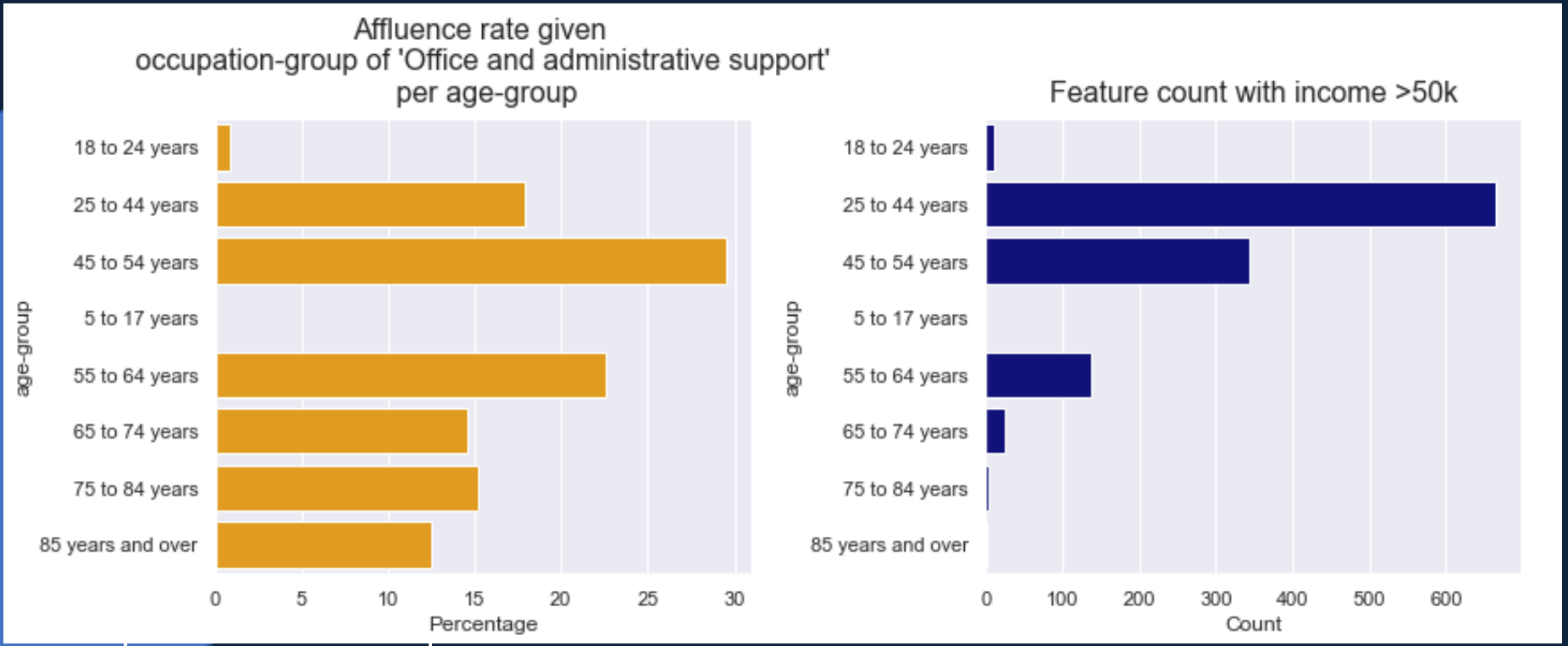
Bivariate Analysis: Occupation vs Age

Regardless of occupation, the probability peaks in the *45 to 54 years* old age group.



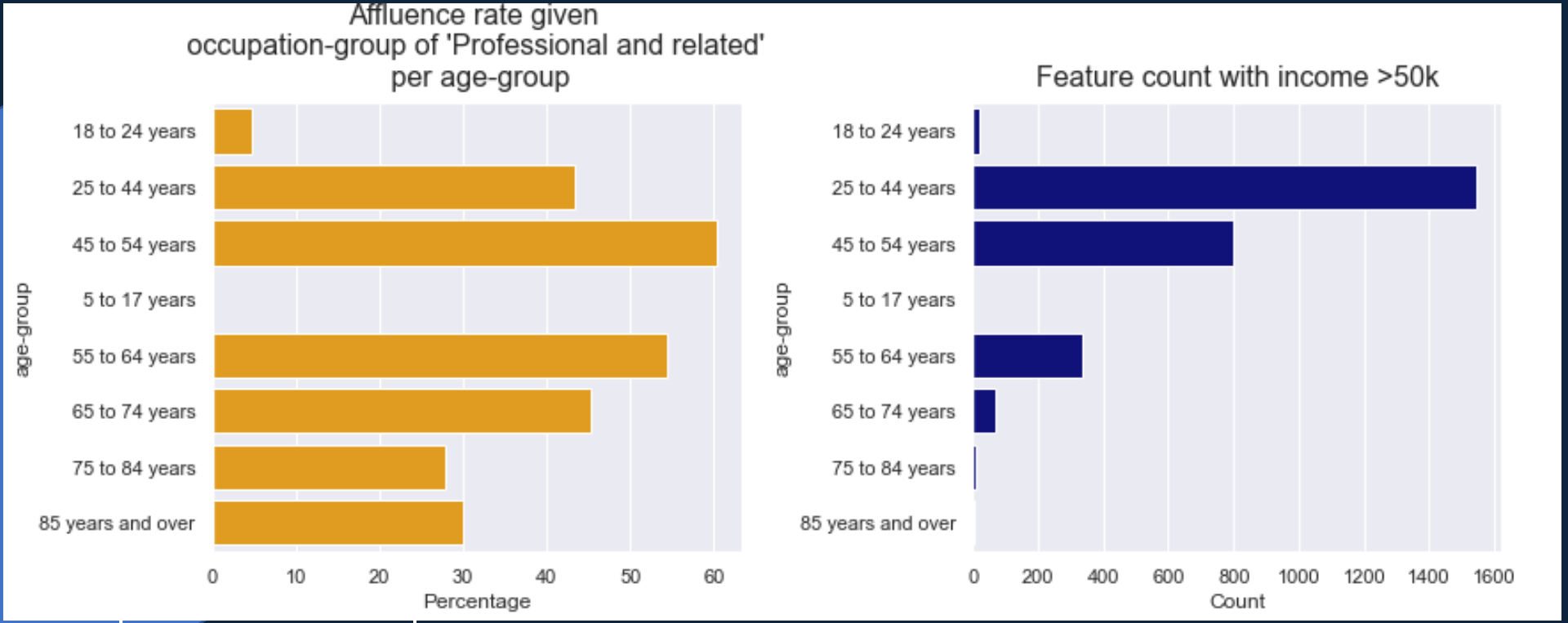
Bivariate Analysis: Occupation vs Age

Regardless of occupation, the probability peaks in the *45 to 54 years* old age group.



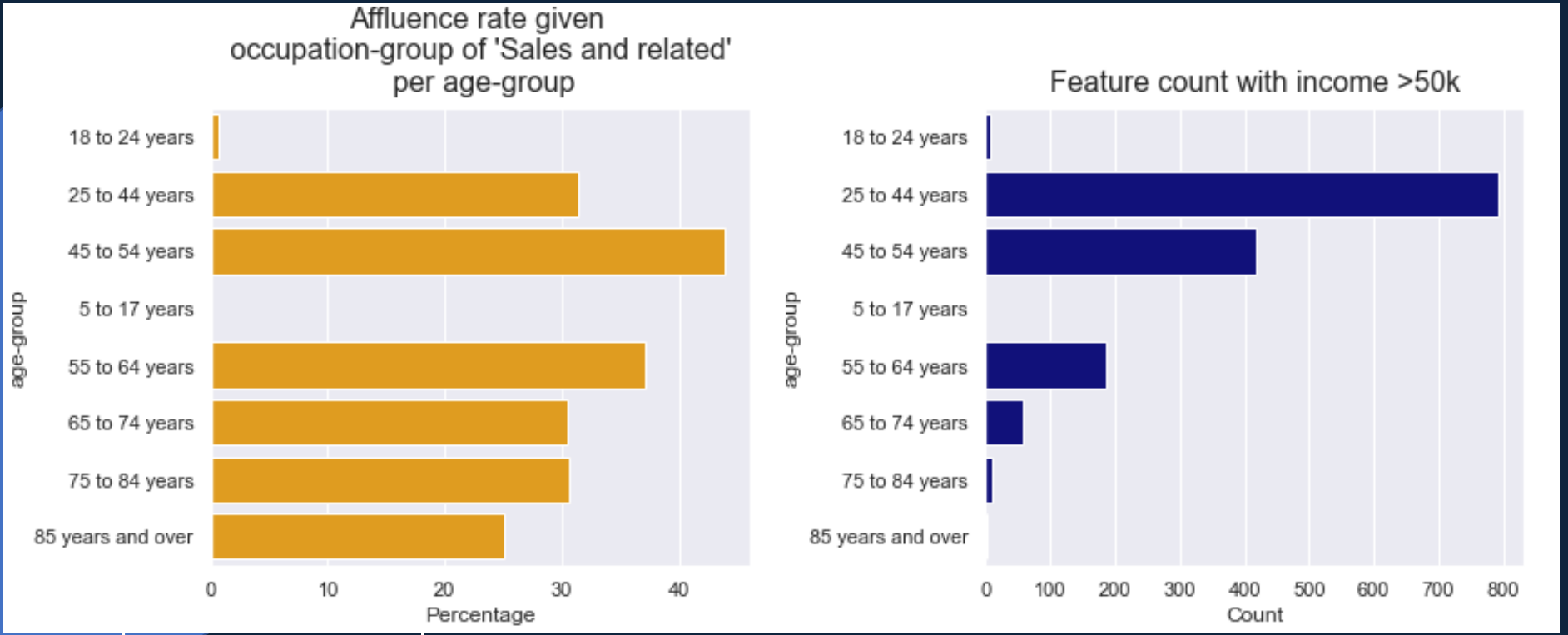
Bivariate Analysis: Occupation vs Age

Regardless of occupation, the probability peaks in the *45 to 54 years* old age group.



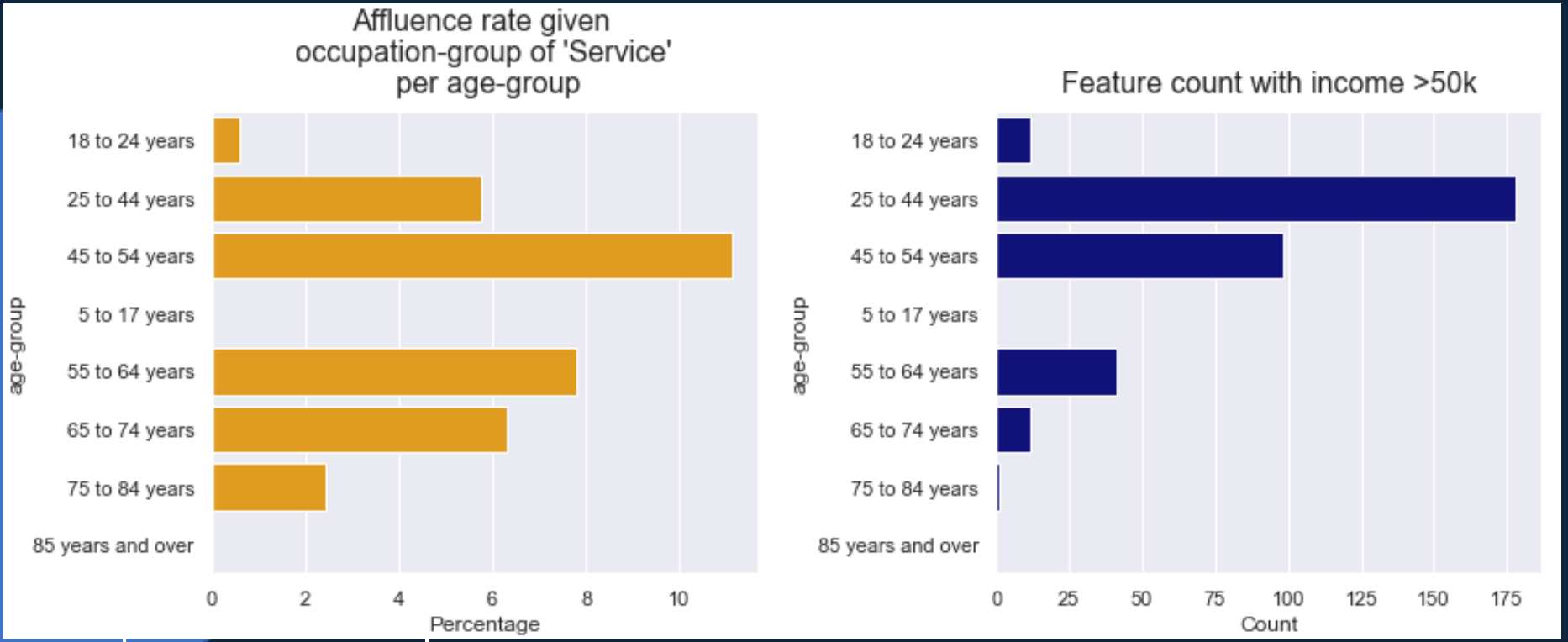
Bivariate Analysis: Occupation vs Age

Regardless of occupation, the probability peaks in the *45 to 54 years* old age group.



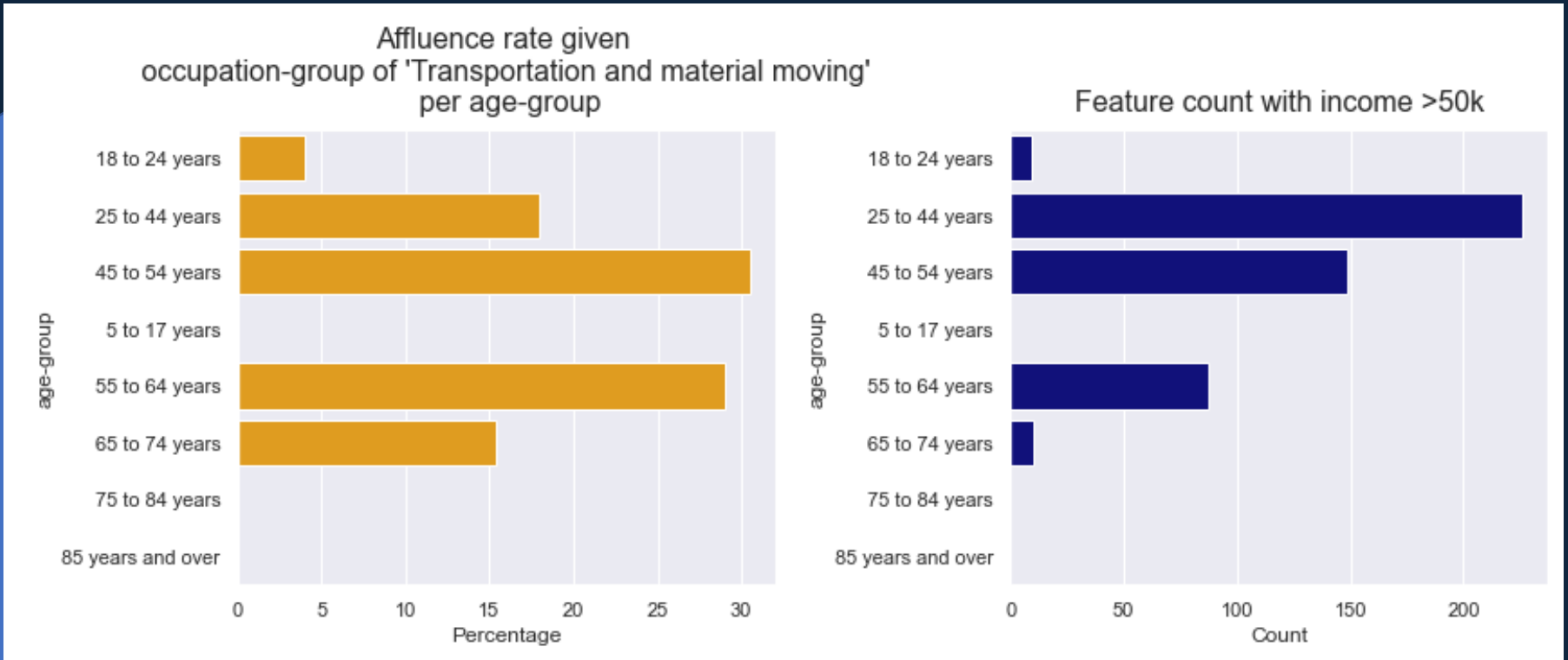
Bivariate Analysis: Occupation vs Age

Regardless of occupation, the probability peaks in the *45 to 54 years* old age group.



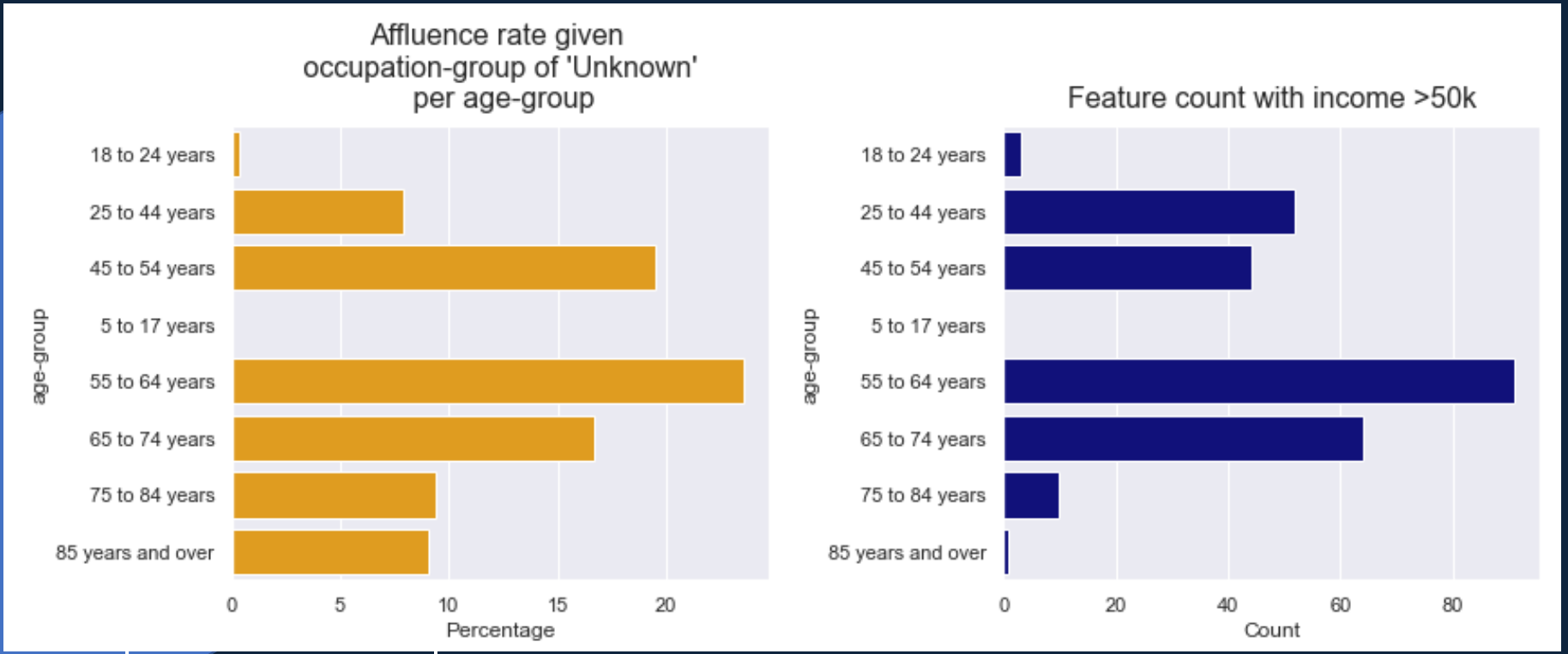
Bivariate Analysis: Occupation vs Age

Regardless of occupation, the probability peaks in the *45 to 54 years* old age group.



Bivariate Analysis: Occupation vs Age

Regardless of occupation, the probability peaks in the *45 to 54 years* old age group.



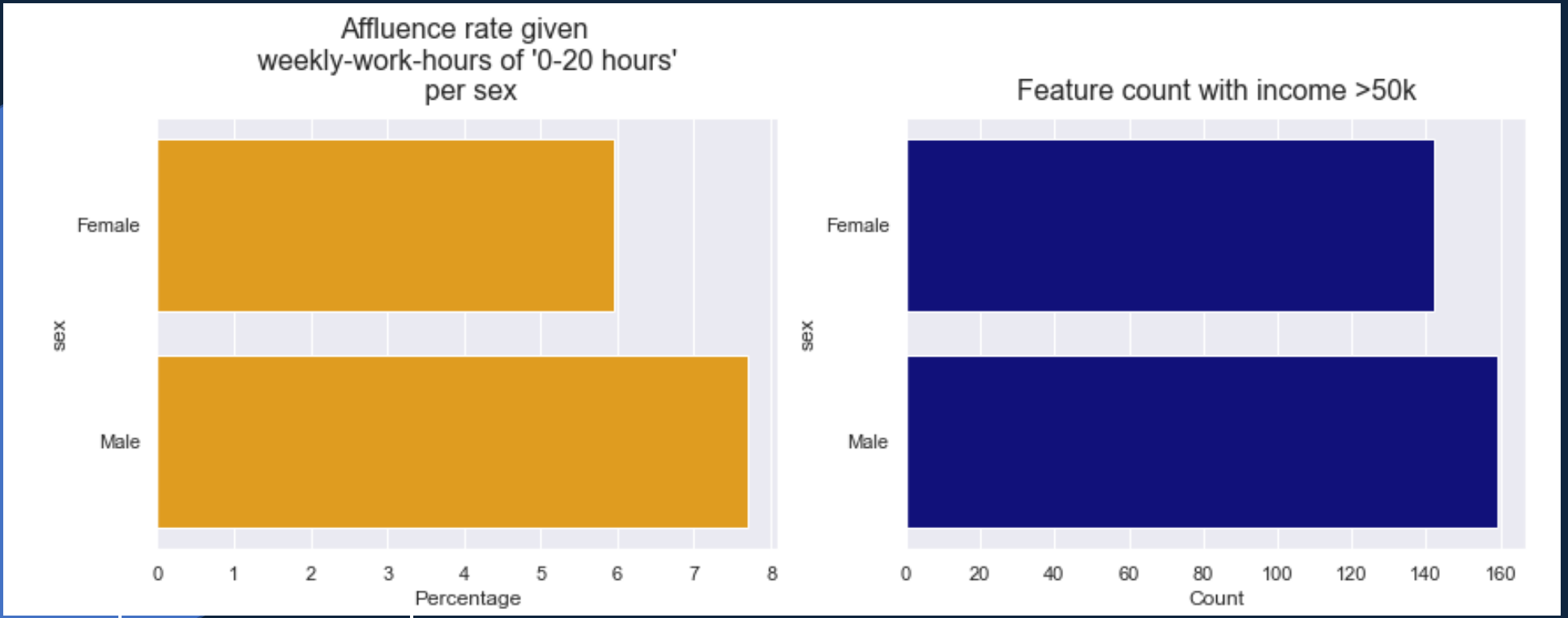
Work Hours vs Sex

- 0 - 20 hours
- 20 – 40 hours
- 40 to 60 hours
- >60 hours

- Female
- Male

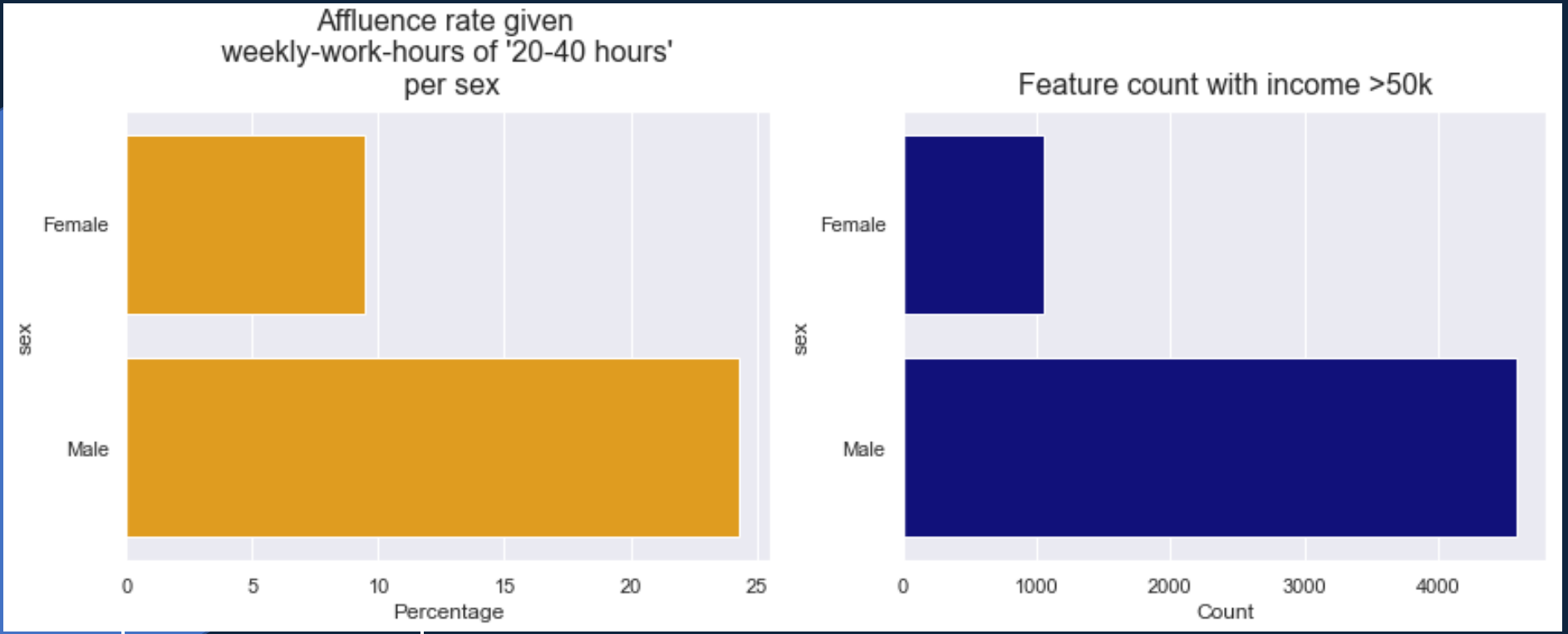
Bivariate Analysis: Work hours vs Sex

Males working for more than *20 hours* have almost twice the probability of having an income >50k than *females*.



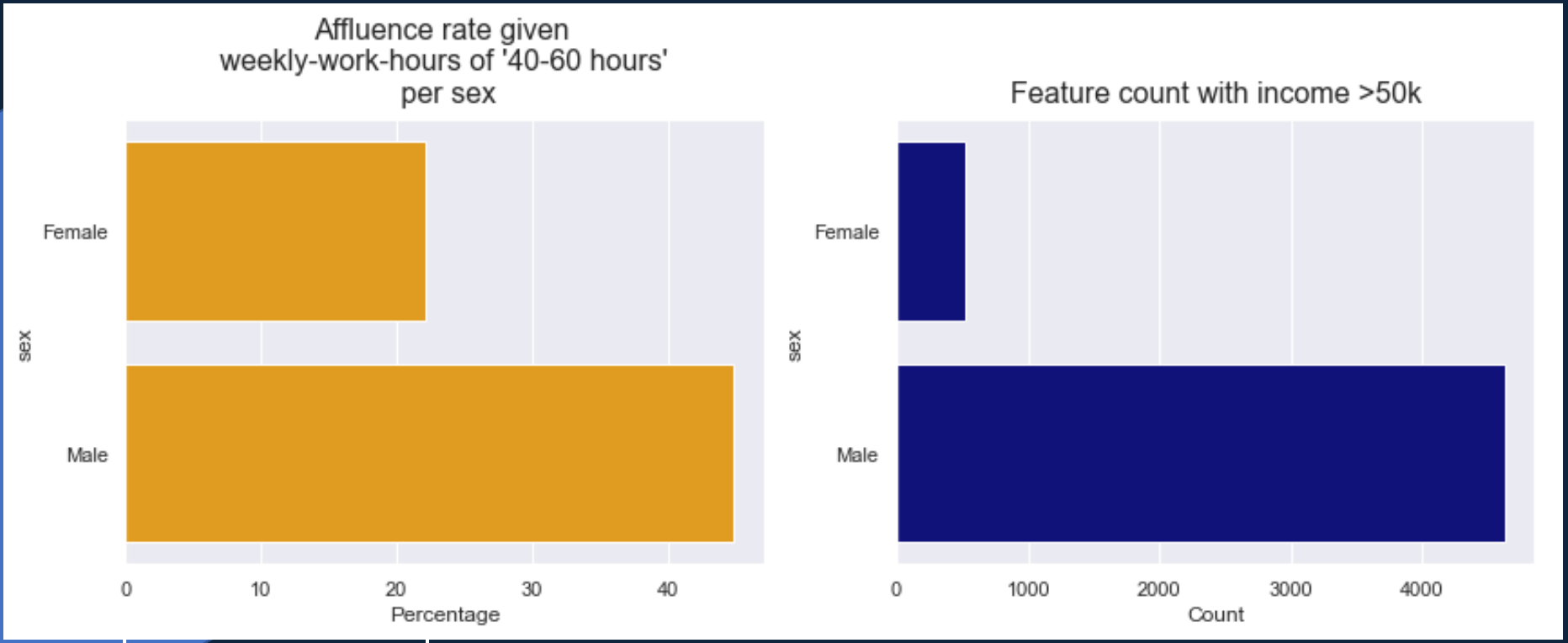
Bivariate Analysis: Work hours vs Sex

Males working for more than *20 hours* have almost twice the probability of having an income >50k than *females*.



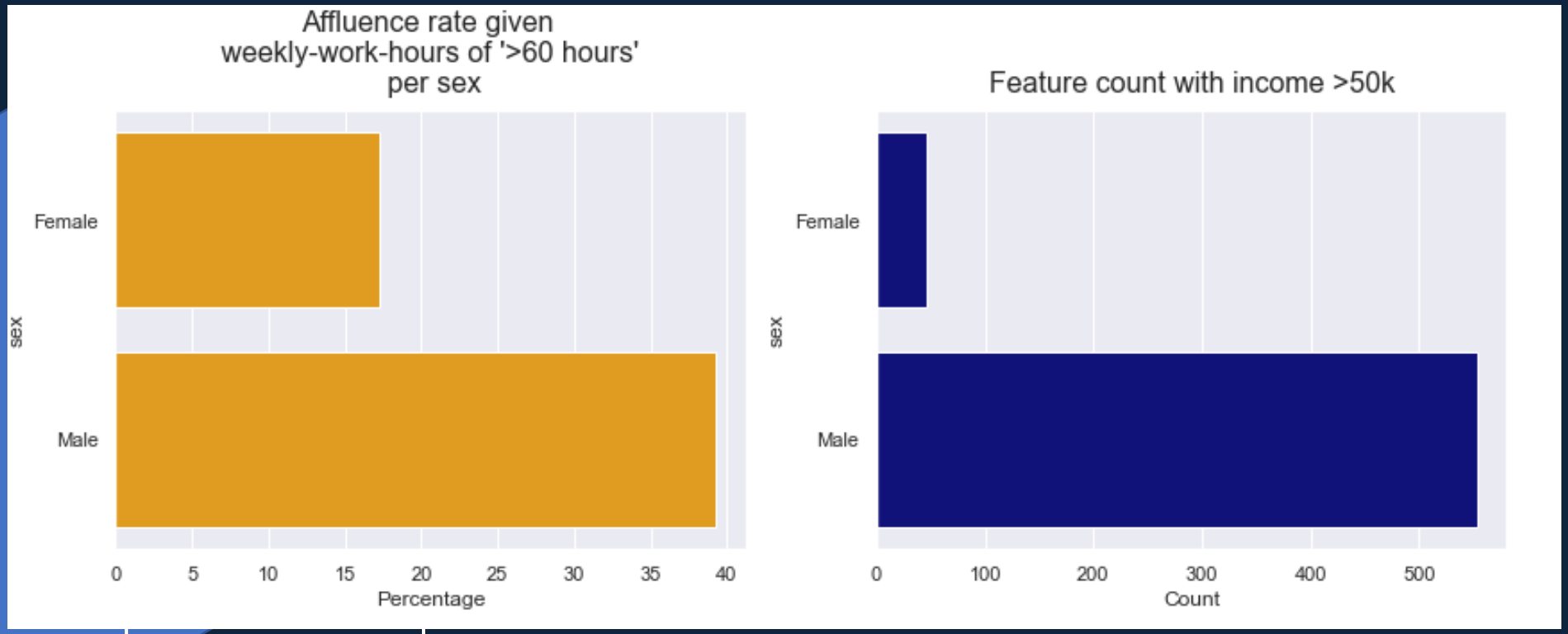
Bivariate Analysis: Work hours vs Sex

Males working for more than *20 hours* have almost twice the probability of having an income >50k than *females*.



Bivariate Analysis: Work hours vs Sex

Males working for more than *20 hours* have almost twice the probability of having an income >50k than *females*.



Summary of Insights

- Working for *60 hours* with *post-graduate degrees* is highly favored (73%).
- *Management, professionals,* and *sales* are the top occupations for bachelor's and post-graduate degree holders.
- *Management, maintenance and repair,* and *transportation and material moving* are the top occupations for non-degree holders.
- Being *self-employed* gives the highest likelihood of having an income >50k among all education groups.
- Age groups between *45 to 54 years* and *55 to 64 years* that are degree holders are more likely to have an income >50k.

Summary of Insights

- *Males* are more likely to have an income >50k than *females, g*iven the same educational attainment.
- The highest-paid jobs in the government are in *management, professionals,* and being in the *military.*
- In comparing *government* and *self-employment*, working *20-40 hours* benefits the self-employed, but the dynamic reverses for *40-60 hours*.
- For *married* and *never-married* people, *male* and *female have equal chances.*
- *Construction*, *military*, *office and adm. support* and *transportation and material moving* are the only jobs that have a higher chance for >60 hours.

Summary of Insights

- Regardless of occupation, affluence rate peaks in the *45 to 54 years* old age group.
- *Males* working for more than *20 hours* have almost twice the probability of having an income >50k than *females*.

ML Models and Review of Related Literature

Review of Related Literature

- [1] A Statistical Approach to Adult Census Income Level Prediction

  Chakrabarty et. al., Oct. 2018 DOI:10.1109/ICACCCN.2018.8748528

- [2] Predicting Annual Income of Individuals using Classification Techniques

  Mohanty et. al., May 2023 DOI: 10.13140/RG.2.2.33330.99529/1

- [3] Classification of Adult Income Using Decision Tree

  Fiagbe, R., "Classification of Adult Income Using Decision Tree" (2023). Data Science and Data Mining. 3. https://stars.library.ucf.edu/data-science-mining/3

Review of Related Literature

- [4] A Comparison of Supervised Learning Algorithms for the Income Classification

  Temraz, International Journal of Computer Applications (0975 – 8887) Volume 182 – No. 38, January 2019

- [5] A Comparative Study of Classification Techniques On Adult Data Set

  Deepajothi et. al., International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 8, October - 2012 ISSN: 2278-0181

- [6] Performance Assessment of Feature Selection methods using K-means on Adult Dataset

  R. Sabitha et. al., Research Journal of Computer Systems Engineering – RJCSE ISSN: 2230-8563; e-ISSN-2230-8571

Classification Models

- Gradient Boosting
- Decision Tree
- Random Forest
- Naïve Bayes
- Logistic Regression
- Support Vector Machine
- Artificial Neural Network

Pre-modeling

Data Reduction

- Features with least Extra Trees Classifier Scores have been dropped: race, native-country

Categorical Data Handling

- LABEL ENCODING – all categorical features are label encoded, where alphabetically each category is assigned numbers starting from 0 up to the number of classes minus 1
- ONE-HOT ENCODING – splitting of different categorical features into its own categories where each and every category assumes a binary value
- DISCRETIZATION – grouping data in a logical way. Trade off is prevention of data overfitting at the cost of granularity

Shuffling

- Shuffle data set in a consistent way such that all the categories of different attributes remain included in training and validation set

Train-Test Split

- Common splits seen are usually 80/20 or 70/30

Modeling and Evaluation

Hyper Parameter Tuning

- Using Grid-Search to find the combination of multiple optimal hyperparameters

Model Evaluation

- TRAINING ACCURACY : how well the model performed on training data
- PRECISION : (also called positive predictive value)

$$\frac{True\ Positives}{(True\ Positives\ +\ False\ Positives)}$$

- RECALL : (also known as sensitivity) given:

$$\frac{True\ Positives}{(True\ Positives\ +\ False\ Negatives)}$$

- F1-SCORE : the harmonic mean of precision and recall

Interpretation

- Important features score and model interpretation

Model Evaluation

		Training Accuracy	F1-score	Remarks
[1]	Gradient Boosting	88.73%	88%	• With hyperparameter tuning • No imbalanced data handling • No out-of-sample accuracy • Features importance and interpretation lacking
[2]	ANN	86.1%	66.60%	• With hyperparameter tuning • For Decision Tree and Random Forest: applied discretization • No imbalanced data handling • No out-of-sample accuracy • Features importance and interpretation lacking
	Decision Tree	85.60% (max depth: 9)	67.54%	
	Random Forest	85.59%	67.86%	
	Naïve Bayes	80.30%	42.8%	
	Logistic Regression	82.60%	55.52%	
	SVM	79.71%	63.68%	
[3]	Decision Tree	85.85%	67.98%	• With hyperparameter tuning • No imbalanced data handling • Important features: *relationship, education-num, capital-gain*

Take aways

Top Models

- Gradient Boosting, ANN, Decision Trees
- Grain of salt for replicable codes. Not all papers provided codes and exact steps to replicate results may not be possible
- Without out-of-sample accuracy, we don't have a way to gauge the confidence we'll give the model

Interpretability

- Papers considered variables that lacked context or definition (ie. essence of capital-gain?)

Hyperparameter Tuning

- Almost indispensable to improve model accuracy

Future works

- Check model performance on group's EDA
- Look into other Encoding Methods
- Look into Sampling Methods (over or under sampling)
- Find complete definitions of variables for interpretability
- Test out-of-sample performance
- Hyper Parameter Tuning

Supplementary

Using the features processed during the group's EDA, we aim to find if the top model from RRL's give us the same level of accuracy with interpretable results

Gradient Boosting

- Full shape: (48842, 9)
- Keep 80% for train-test, 20% for out-of-sample

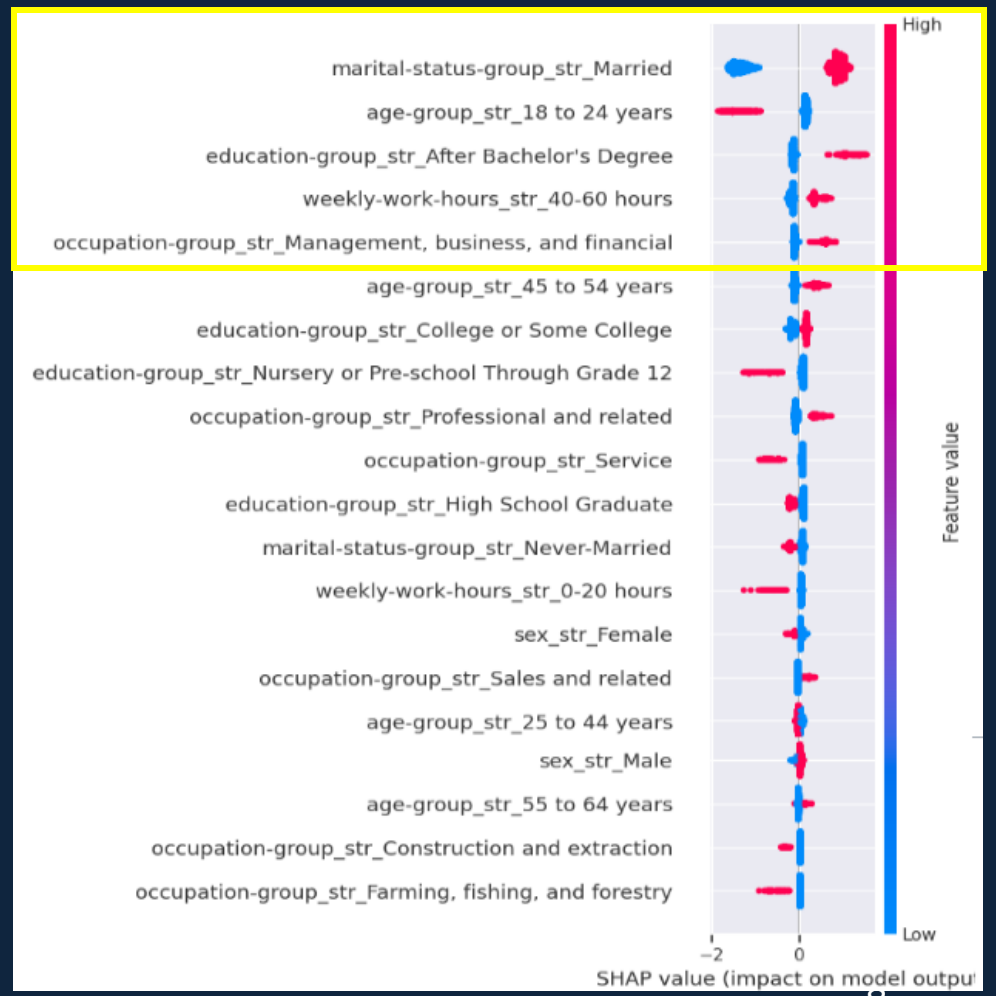
Pre-modeling:

- Data reduction
- Data discretization
- Down sampling: income in {0, 1} : (9376, 9)
- One-hot-encoding for categorical data
- No shuffling
- No hyperparameter tuning

Training Result:

- Training Accuracy, F-1 Score: 80.44% and 80.38%
- Out-of-sample Accuracy, F1-score: 77.71% and 74.15%

Important Features



Thank you

Cesar Malenab

Regina Flores

Emmanuel Pedernal