

Figure 1. Income by industry and occupation. Same occupation in different industries yields different income.

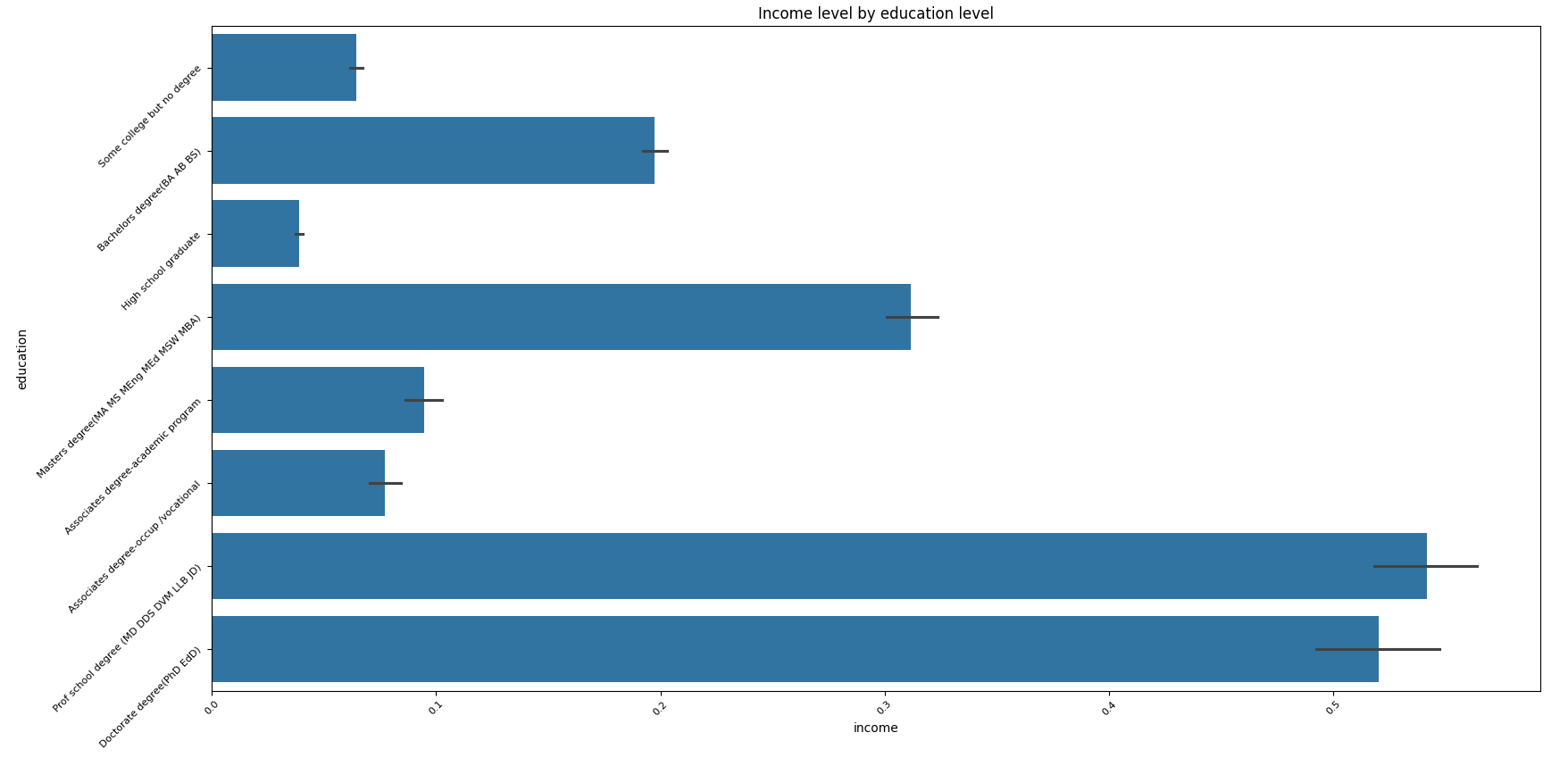


Figure 2. Income by education level. The more you formally study the higher the chance of earning > 50k.

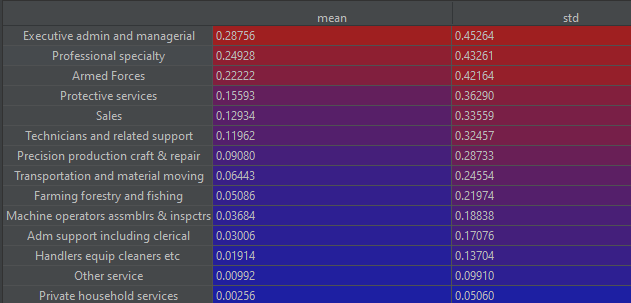


Figure 3. Top occupations by chance of getting paid more than 50k.



Figure 4. Top industries by chance of getting paid > 50k.

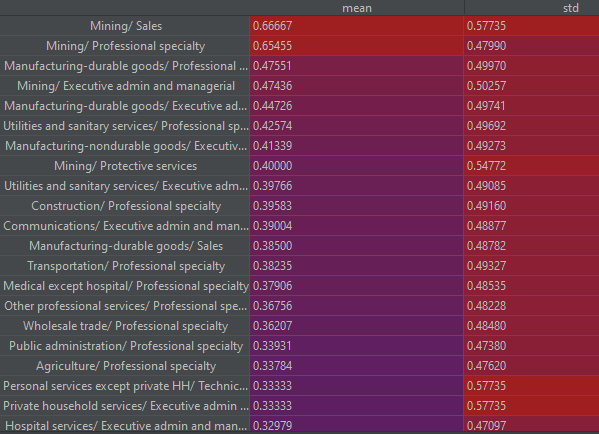


Figure 5. Highest chance of high salary (>50k) by an industry/occupation combination.

In the following code snippets I explain why I decided to remove some of the columns in the data set.

## The data can be simplified in some of the columns

# train[train.ACLSWKR.isin([' Never worked', ' Without pay'])]['income'].value\_counts()

# income

# 0 601

# 1 3

# train[train.AHGA==' Children']['income'].value\_counts()

# income

# 0 47422

# train.groupby('AHGA')['income'].agg(['mean', 'std']).sort\_values(by='mean', ascending=False)

# mean std

# AHGA

# Prof school degree (MD DDS DVM LLB JD) 0.540435 0.498501

# Doctorate degree(PhD EdD) 0.520190 0.499790

# Masters degree(MA MS MEng MEd MSW MBA) 0.311573 0.463172

# Bachelors degree(BA AB BS) 0.197080 0.397803

# Associates degree-academic program 0.094430 0.292460

# Associates degree-occup /vocational 0.077081 0.266745

# Some college but no degree 0.064234 0.245174

# High school graduate 0.038818 0.193162

# 12th grade no diploma 0.015992 0.125476

# 11th grade 0.010180 0.100390

# 7th and 8th grade 0.008992 0.094405

# 10th grade 0.008204 0.090211

# 1st 2nd 3rd or 4th grade 0.007226 0.084723

# 5th or 6th grade 0.006713 0.081673

# 9th grade 0.006100 0.077867

# Less than 1st grade 0.001221 0.034943

# Children 0.000000 0.000000

# train.AHSCOL.value\_counts(dropna=False)

# AHSCOL

# NaN 186942

# High school 6892

# College or university 5688

# train.AMARITL.value\_counts()

# AMARITL

# Never married 86485

# Married-civilian spouse present 84222

# Divorced 12710

# Widowed 10462

# Separated 3460

# Married-spouse absent 1518

# Married-A F spouse present 665

# drop columns

# train.AREORGN.value\_counts(dropna=False)

# AREORGN

# All other 171906

# Mexican-American 8079

# Mexican (Mexicano) 7234

# Central or South American 3895

# Puerto Rican 3313

# Other Spanish 2485

# Cuban 1126

# NA 874

# Do not know 306

# Chicano 304

# train.groupby('HHDREL')['income'].mean()

# HHDREL

# Child 18 or older 0.008732

# Child under 18 ever married 0.000000

# Child under 18 never married 0.000040

# Group Quarters- Secondary individual 0.007576

# Householder 0.127870

# Nonrelative of householder 0.030785

# Other relative of householder 0.008864

# Spouse of householder 0.054712

# train.groupby('ARACE')['income'].mean().sort\_values(ascending=False)

# ARACE

# Asian or Pacific Islander 0.073693

# White 0.067350

# Black 0.026451

# Other 0.024884

# Amer Indian Aleut or Eskimo 0.021768

# train.groupby('same\_house')['MARSUPWT'].sum()

# same\_house

# 0.0 2.862201e+07

# 1.0 1.412779e+08

# train.groupby('PARENT')['MARSUPWT'].sum().sort\_values(ascending=False)

# PARENT

# Both parents present 64329131.18

# Mother only present 22027634.35

# Father only present 3151752.83

# Neither parent present 3008288.05

# train.groupby('PARENT')['MARSUPWT'].mean()

# PARENT

# Both parents present 1650.184213

# Father only present 1673.793324

# Mother only present 1724.681675

# Neither parent present 1819.895977