CSE 578 Final Report

By: Kexin Tong

Goals and Business Objectives

The goal of this report is to determine a demographic profile for UVW to market its degree programs based on criteria of targeting individuals making more and less than $50,000 in salary.

Dataset from the United States Census Bureau will be used as our dataset for this analysis. We will identify important factors that contribute to a person's income and analyze whether income can be predicted given a person's profile.

This report will highlight several key features that contribute to income and dive into some patterns and visualizations accompanying each.

Assumptions

Assumption 1: Data provided is measured accurately. We assume the data is free of errors and will not mislead. We assume all the factors provided by each individual in this dataset are correct and true.

Assumption 2: Data provided is sampled without gaps and without bias. We assume the data is a fair and complete representation of the geographic area that it covers. For example, if the samples are taken only from hospital workers and their families, it would skew the results compared to if we took a consensus from a whole state.

Assumption 3: Data provided is relevant and timely. We assume the data collected is from a recent and that the results can be extrapolated to current time. If the data collected is too far in the past, it could be irrelevant or misrepresent when we apply it to the present. For example, pre industrial age or recession times may show a very different set of data.

Assumption 4: I assume the results we draw from the important features would be a driving factor in determining income. Since we are analyzing and producing visualization on only these features in the dataset, we assume that these features will be a strong contributor to income.

User Stories

Below are the five user stories we defined and will analyze in this report.

User Story 1: I want to know if age is a reliable indicator of determining income.

User Story 2: I want to know if education is a reliable indicator of determining income.

User Story 3: I want to know if marital status, sex, relationship and occupation are reliable indicators in determining income.

User Story 4: I want to know if there are any meaningful relationships between capital gain, capital loss and salary.

User Story 5: I want to find out if there are any meaningful relationships between age, education, capital gains and salary.

Visualizations

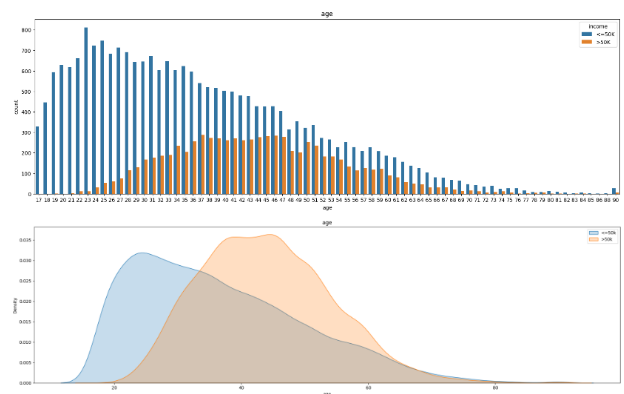
I started off importing the data into a dataframe, adding the column names, and cleaning the data of any rows with '?' variables by removing these rows from the dataframe. I then converted the string columns into categories so they can eventually be mapped into numbers for the prediction models in future steps. Now our cleaned data frame has 22k rows with >$50k income and 7k rows with <$50k income.

I built a function to summarize numeric variables by looking at its breakdown of the number of variables with >$50k and <$50k income and finding out the min, max, average, standard deviation of total and each income category. I also added a histogram plot to see if the distribution of salary varies greatly for the particular variable, and added a box plot to see what the median and outliers look like. I ran this single variate analysis function through all the numeric variables and found that education, marital status, occupation and relationship may be important factors in determining incomes.

Next, I built a function to evaluate categorical variables by summing the number of datapoints for each category, along with the breakdown into >$50k income and <$50k income. I also added a pie graph to visualize total data points per category, a histogram for category count by income and a mosaic plot to see the breakdown of each category by income. I ran this single variate analysis function through all the categorical variables and found that age. capital_gains and education_num are important factors in determining incomes, while the others do not look to be a meaningful factor in determining income.

Lastly, I built a function for multivariate analysis between two or more numeric variables by plotting the scatter matrix of each variable against each other, colored by income. If there are any two variables with two distinct clusters of the two income levels this could identify a strong candidate profile of income level.

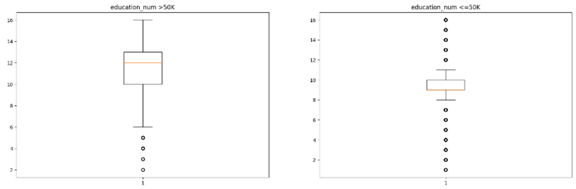
User Story 1 - Age
Age is a meaningful factor in determining income as we see that as age increases salary decreases. There is a higher chance of making >$50k salary for a person aged between late 30s to early 50s, while a person aged between early 20s to early 30s is likely to make <$50k salary.



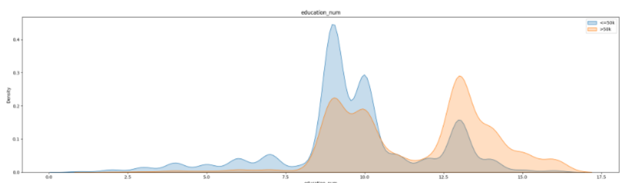
I chose to use a histogram because plotting both salary amounts in one plot can clearly show whether they are similar or not, and show the difference in average, median and how the two series differ. From here, we clearly see that $>50k and $<50k have very different distributions.

I then chose to use a kernel density plot to look closer at the density of both incomes. Here, we can immediately tell that a younger age range has a higher probability of making $<50k while an older age range is much more likely to make $>50k.

User Story 2 - Education
Education is a meaningful factor in determining income. We can see from below that for the factor education, having some college degree (education_num 10) or above have a significantly higher probability of making a salary >$50k.



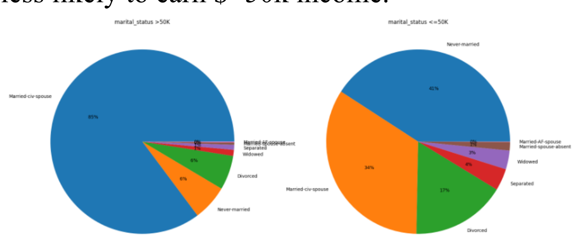
We can also see this reflected in the density chart using the education_num factor, which shows that for 10 (some college) and above, there is more chance of making >$50k salary.

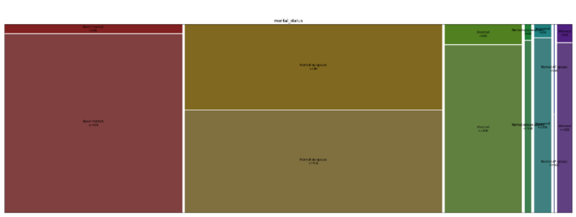


I chose to use box plots because we immediately see five number statistics outlined, and how different they are for both incomes. Immediately, we see the median for $>50k income is much higher, and the quantiles also cover a higher range than the $<50k group.

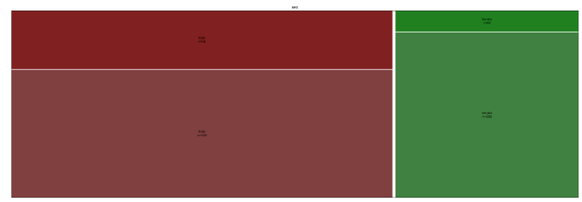
User Story 3 - Marital status, Sex and Occupation

Marital status is an important contributor as we see that this dataset is highly skewed to show that married spouses are much more likely to make higher income, while never married are less likely to earn $>50k income.

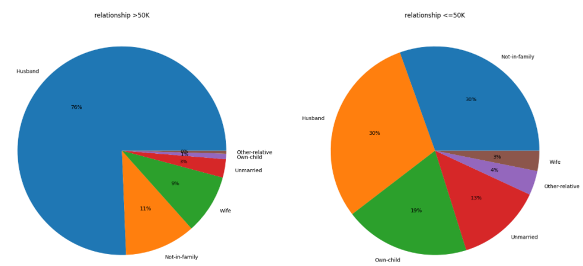


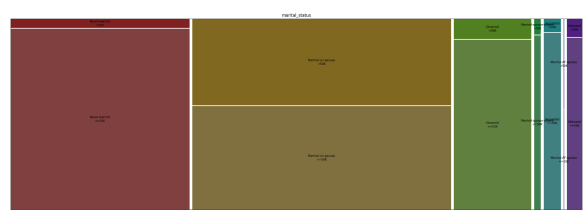


Sex may not be a strong indicator of income because we see that while both sexes have higher income and there are more male than females, there's no notable skew in the income distribution by sex.

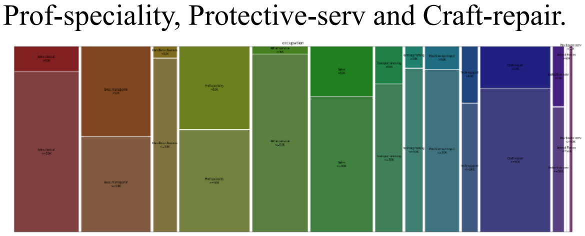


Relationships can also be an important contributor to income as we see that husband and wife have more chance of making higher income, which ties into the insight we drew from the marital status factor.





Occupation is another important contributor to income, as we see distribution between the two classes are highly skewed. Some occupations more likely to make $>50k are Exec-managerial, Prof-speciality, Protective-serv and Craft-repair.

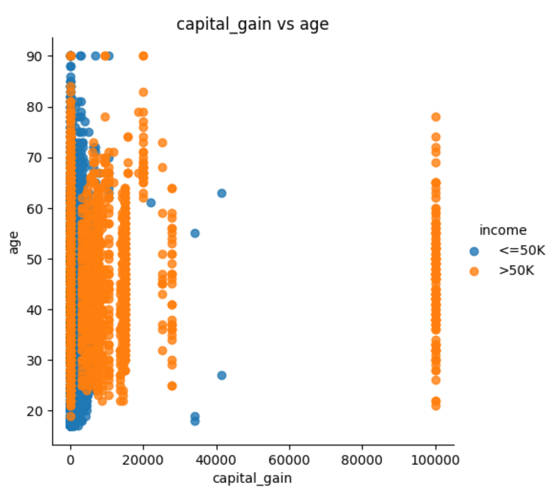


I used pie charts as the visualization to show the breakdown of two income levels by category, as the number of data points in the $<50k group is twice as large as the data points in the $>50k group and pie charts present percentages. For example, looking at the marital status pie charts I can immediately tell that most married spouses are likely to make higher income because their percentage is much higher than any other category in this pie chart.

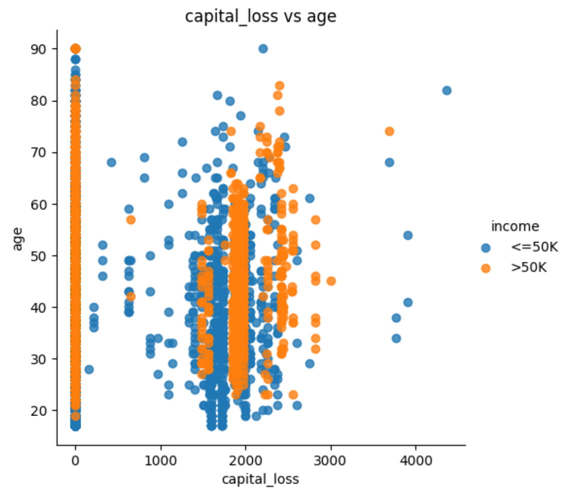
I chose to use mosaic plots for single variate categorical analysis because they best show the breakdown of each category and their percentage of data points for each income. Since we have more data points for the lower income group, if there is a particular category with the break of higher vs lower income in the middle or lower, it could be a sign that this category may be a significant contributor.

User Story 4 - Relationship between Capital Gains and Capital Loss vs Income

We see that users across all ages with a high capital gains are likely to make a salary >$50k.



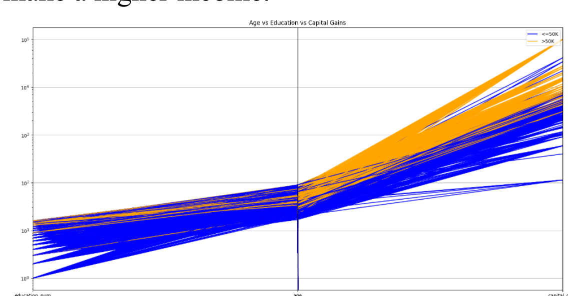
While for capital loss, the isn't enough skew to make a conclusion that it is meaningfully different for $>50k and $<50k income.



I chose to use scatter plots because we immediately see the difference in capital gains vs capital loss's distributions.

User Story 5 - Relationship between Age, Education, Capital Gains and Salary

Here we see a relationship between age, education and salary. Generally, we see that higher age with higher capital gains are likely to make a higher income.



I chose to use a parallel coordinate plot because it is a useful visualization to compare multiple variables together and see if there is a relationship between them. Here we can see that the orange line and blue lines can be distinguished using these three features.

Questions and Solutions

Question 1: Which visualizations are best for us to use?

For numeric variables, the visualizations used were histogram charts, density plots, box plots and scatterplots. For categorical variables we chose histogram plots, pie charts and mosaic charts. Then we tested out every variable in each visualization to pick the ones more representative of each to draw conclusions on.

Question 2: How to compare categorical data vs numerical data?

In order to compare categorical data to numerical data, we first needed to convert the categorical variables into numeric. This was done by converting each column into an associate number just like education and education_num provided.

Question 3: How to find a suitable metric to determine if a categorical variable is important for determining salary?

After plotting pie graphs, mosaic plots, histograms, converting each category into a numerical number and running the numeric analysis function, I realized that histogram and mosaic plots are most useful because it is easy to tell whether breakdown of income differs from category to category, and thus I chose to use these plots in my visualization.

Future Plan

Beyond the scope of this report, I want to build a prediction model that tests 2-3 machine learning models using the important features I identified above to find an algorithm that predicts the salary amount of the testing data. The algorithms I want to test will be a subset of decision tree, k-nearest neighbor, regression and random forest. I will use a 80-20 split for testing and training data to evaluate the performance of each model, using accuracy, precision and recall.

Another future step is to suggest a suitable marketing choice after we have the target demographic. We can analyze which media avenue our target demographic utilizes the most by finding a new dataset and running an analysis. It could be TV commercials to newspaper to facebook ads depending on the target for example.

Lastly we can also include a feasibility study to scope out all the critical aspects of this project and find out its likelihood of successfully onboarding new students to UVW.

Appendix

Full code is attached below.

In [83]:
```
import sqlite3
import pandas as pd
import numpy as np
from matplotlib import pyplot as plt
import seaborn as sns
from pandas.plotting import parallel_coordinates
from statsmodels.graphics.mosaicplot import mosaic
%matplotlib inline
colnames = ['age', 'workclass', 'fnlwgt', 'education', 'education_num', 'marital_st
data = pd.read_csv('Data/adult.data', sep=", ", names=colnames, header=None, index_
data.replace("?", float("nan"), inplace=True)
data.dropna(inplace=True)

# create mappings for education
education_mapping = data[['education', 'education_num']].drop_duplicates().sort_val
print(education_mapping)
```

```
C:\Users\Kathy\AppData\Local\Temp\ipykernel_13024\4085719864.py:10: ParserWarning:
Falling back to the 'python' engine because the 'c' engine does not support regex
separators (separators > 1 char and different from '\s+' are interpreted as rege
x); you can avoid this warning by specifying engine='python'.
  data = pd.read_csv('Data/adult.data', sep=", ", names=colnames, header=None, ind
ex_col=False)
        education  education_num
224     Preschool              1
416       1st-4th              2
56        5th-6th              3
15        7th-8th              4
6             9th              5
219          10th              6
3            11th              7
415          12th              8
2         HS-grad              9
10   Some-college             10
48      Assoc-voc             11
13     Assoc-acdm             12
0       Bachelors             13
5         Masters             14
52    Prof-school             15
20      Doctorate             16
```

In [84]:
```
# clean data/change data types:
data.workclass = data.workclass.astype('category')
data.education = data.education.astype('category')
data.marital_status = data.marital_status.astype('category')
data.occupation = data.occupation.astype('category')
data.relationship = data.relationship.astype('category')
data.race = data.race.astype('category')
data.sex = data.sex.astype('category')
data.native_country = data.native_country.astype('category')
data.income = data.income.astype('category')
```

In [85]:
```
print("Count of income >50k: " + str(len(data[data.income == '>50K'])))
print("Count of income <=50k: " + str(len(data[data.income == '<=50K'])))
```

```
Count of income >50k: 7508
Count of income <=50k: 22654
```

In [86]:
```
data.head()
```

Out[86]:

	age	workclass	fnlwgt	education	education_num	marital_status	occupation	relationship	r
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	Wl
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	Wl
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	Wl
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Bl
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Bl

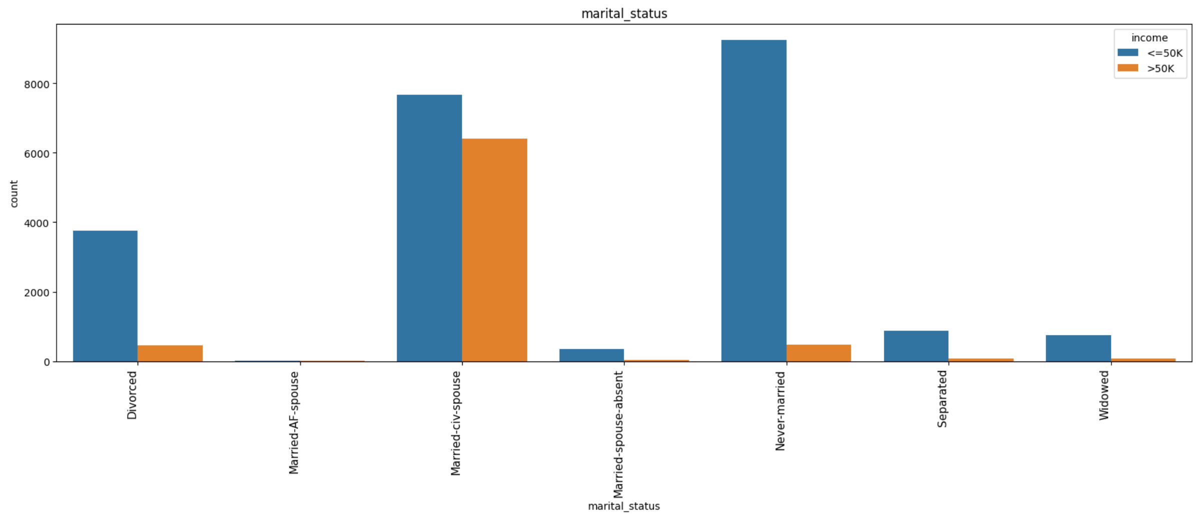
In [87]:
```
# visualization for all single variables:

def numerical(col):
    print("For variable " + col)
    print("Min is: " + str(data[col].min()))
    print("Max is: " + str(data[col].max()))
    print("Mean is: " + str(data[col].mean()))
    print("Std Dev is: " + str(data[col].std()))

    # countplot of all values based on income
    fig1 = plt.figure(figsize=(20,6))
    ax1 = sns.countplot(x=col, hue='income', data=data)
    plt.title(col)

    # kdeplot for all values based on salary
    fig2, ax2 = plt.subplots(figsize=(20, 6))
    sns.kdeplot(data[data.income=="<=50K"][col], label='<=50k', fill=True, ax=ax2)
    sns.kdeplot(data[data.income==">50K"][col], label='>50k', fill=True, ax=ax2)
    ax2.legend()
    plt.title(col)
    plt.tight_layout()

    # scatterplot vs income
    fig3, ax3 = plt.subplots(ncols=2, nrows=1, figsize=(20,6))
    ax3[0].boxplot(data[data.income=='>50K'][col])
    ax3[0].set_title(col+" >50K")
    ax3[1].boxplot(data[data.income=='<=50K'][col])
    ax3[1].set_title(col+" <=50K")
    plt.show()


def categorical(col):
    counts_above = data[data.income=='>50K'][col].value_counts()
    counts_below = data[data.income=='<=50K'][col].value_counts()
```

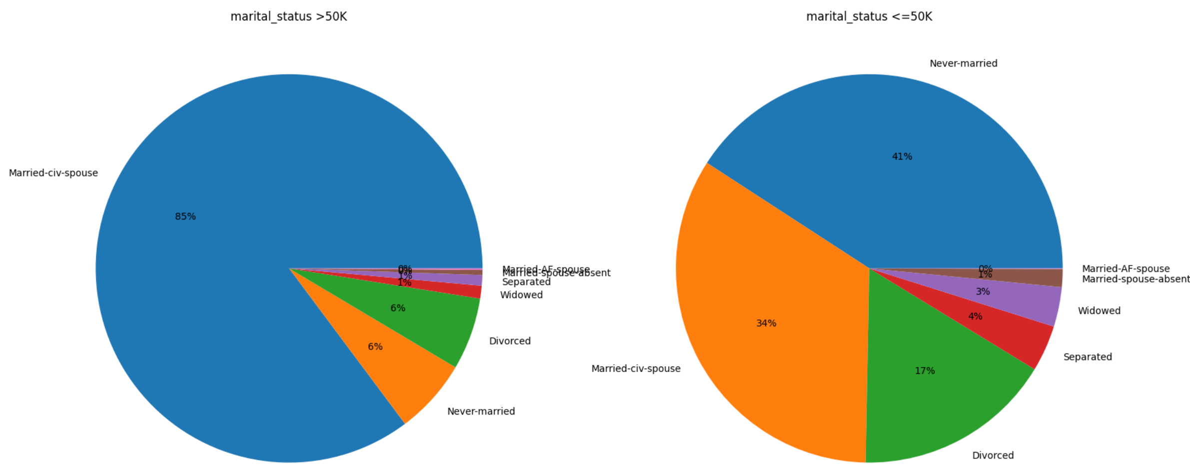
```
    print("For variable " + col)
    print(counts_above)
    print(counts_below)

    # countplot of all values based on income
    fig1 = plt.figure(figsize=(20,6))
    ax1 = sns.countplot(x=col, hue='income', data=data)
    plt.xticks(rotation=90, ha='right', fontsize=11)
    plt.title(col)
    plt.show()

    # pie chart of all values
    fig2, ax2 = plt.subplots(ncols=2, nrows=1, figsize=(20,10))
    ax2[0].pie(counts_above, labels = counts_above.index, autopct='%.0f%%')
    ax2[0].set_title(col+" >50K")
    ax2[1].pie(counts_below, labels = counts_below.index, autopct='%.0f%%')
    ax2[1].set_title(col+" <=50K")
    plt.show()

    # matrix/heatplot vs income
    fig3, ax3 = plt.subplots(figsize=(30,10))
    mosaic(data, [col, 'income'], ax=ax3, axes_label=False)
    plt.title(col)
    plt.show()
```

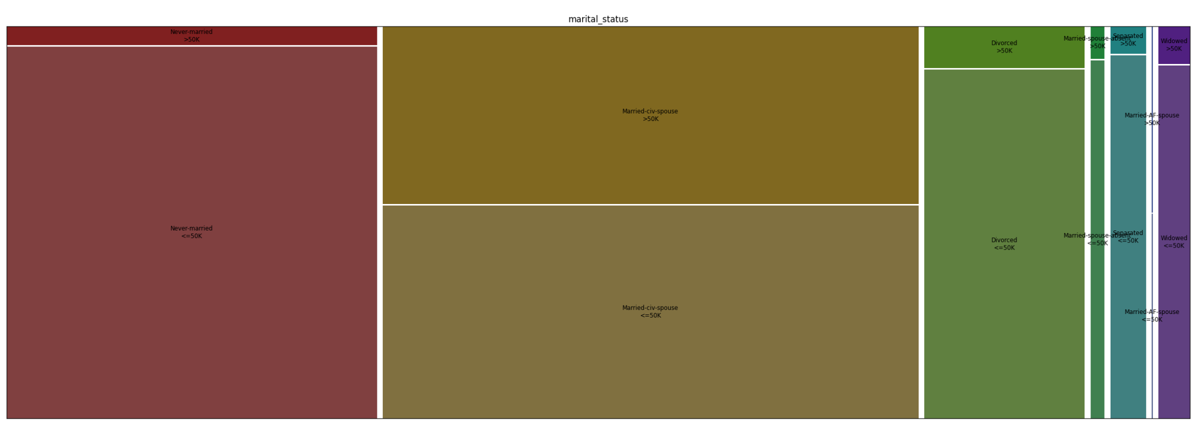
In [88]:

```
categorical('marital_status')
categorical('sex')
categorical('occupation')
categorical('relationship')
```

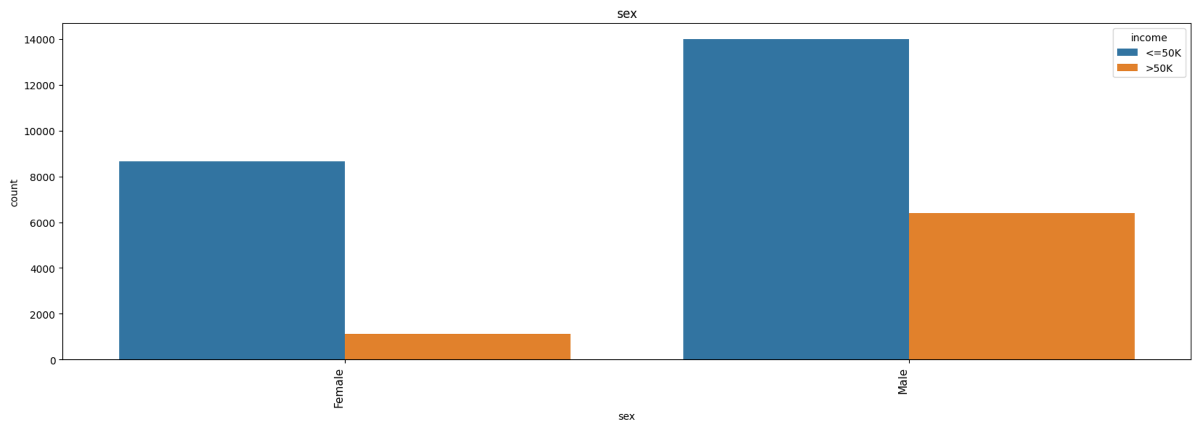
```
For variable marital_status
Married-civ-spouse       6399
Never-married             470
Divorced                  452
Widowed                    80
Separated                  66
Married-spouse-absent      31
Married-AF-spouse          10
Name: marital_status, dtype: int64
Never-married            9256
Married-civ-spouse       7666
Divorced                 3762
Separated                 873
Widowed                   747
Married-spouse-absent     339
Married-AF-spouse          11
Name: marital_status, dtype: int64
```

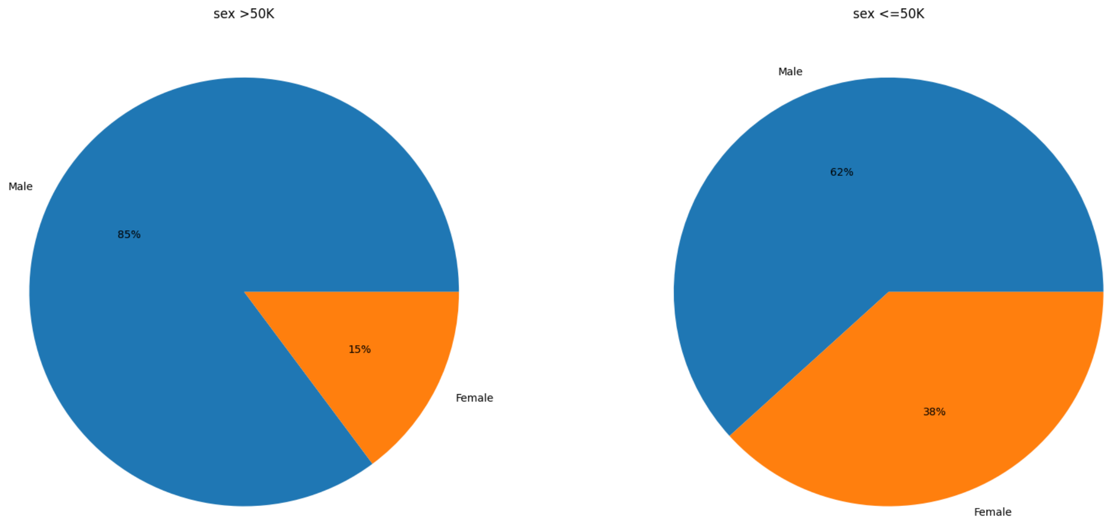


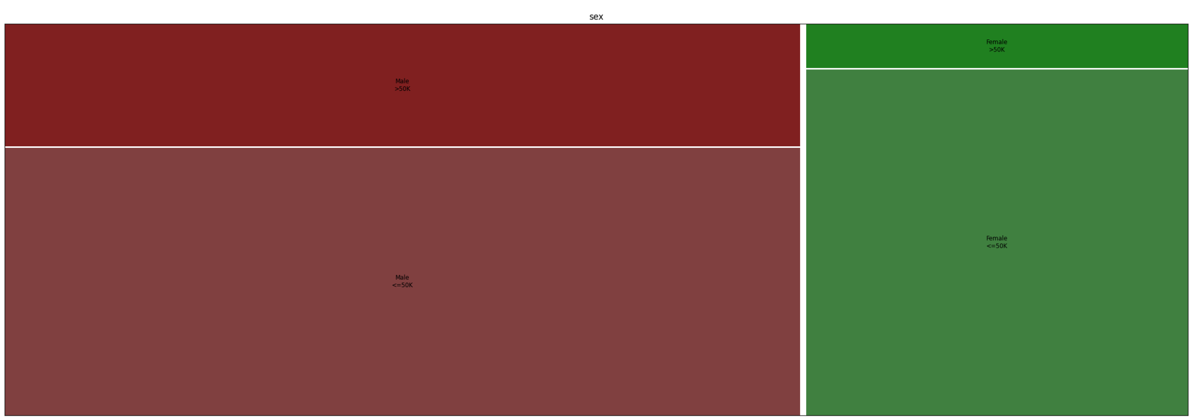




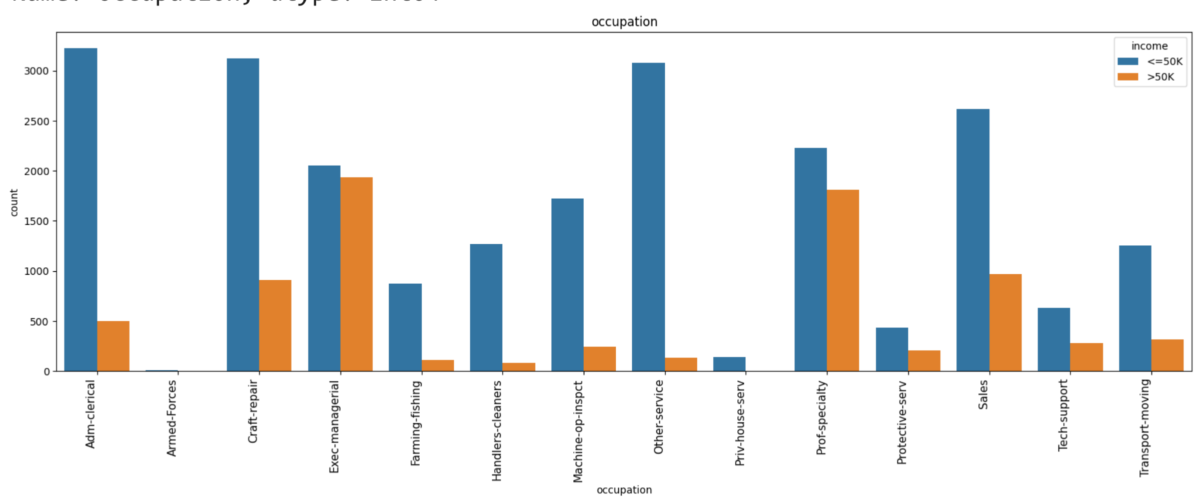

```
For variable sex
Male      6396
Female    1112
Name: sex, dtype: int64
Male      13984
Female     8670
Name: sex, dtype: int64
```

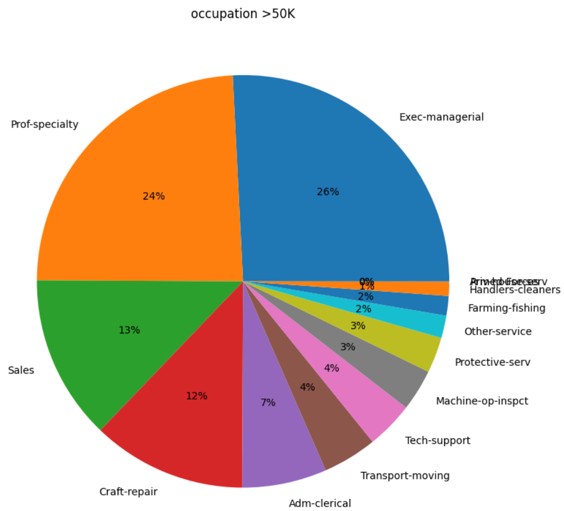

sex
income
<=50K
>50K
count
0
2000
4000
6000
8000
10000
12000
14000
Female
Male
sex

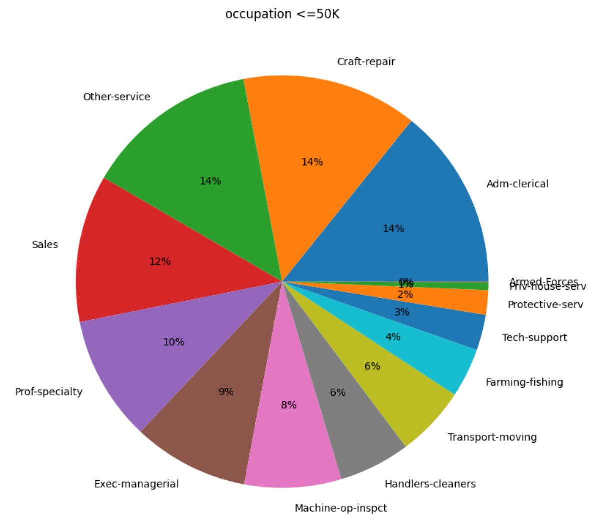

sex >50K
Male
85%
15%
Female
sex <=50K
Male
62%
38%
Female

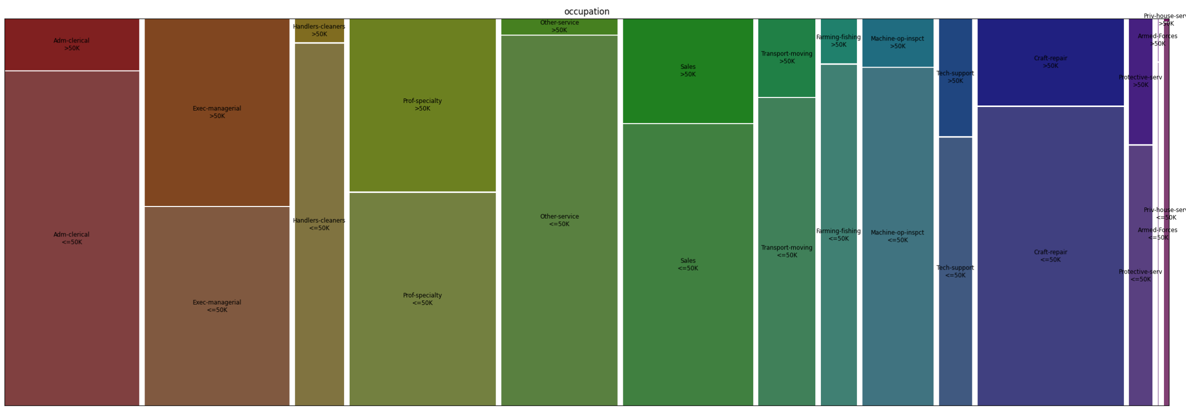

sex
Male
>50K
Male
<=50K
Female
>50K
Female
<=50K

```
For variable occupation
Exec-managerial      1937
Prof-specialty       1811
Sales                 970
Craft-repair          908
Adm-clerical          498
Transport-moving      319
Tech-support          278
Machine-op-inspct     245
Protective-serv       210
Other-service         132
Farming-fishing       115
Handlers-cleaners      83
Armed-Forces            1
Priv-house-serv         1
Name: occupation, dtype: int64
Adm-clerical         3223
Craft-repair         3122
Other-service        3080
Sales                2614
Prof-specialty       2227
Exec-managerial      2055
Machine-op-inspct    1721
Handlers-cleaners    1267
Transport-moving     1253
Farming-fishing       874
Tech-support          634
Protective-serv       434
Priv-house-serv       142
Armed-Forces            8
Name: occupation, dtype: int64
```

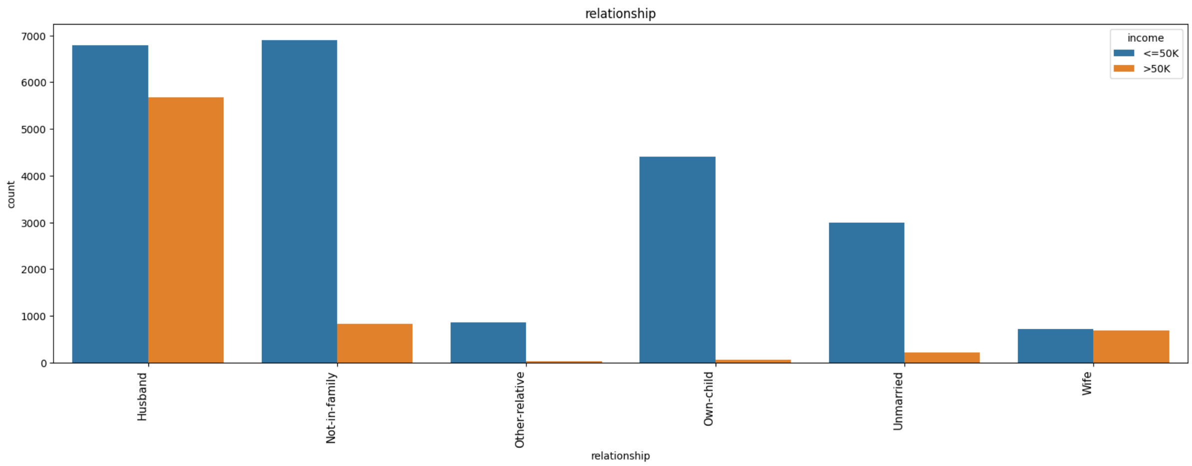


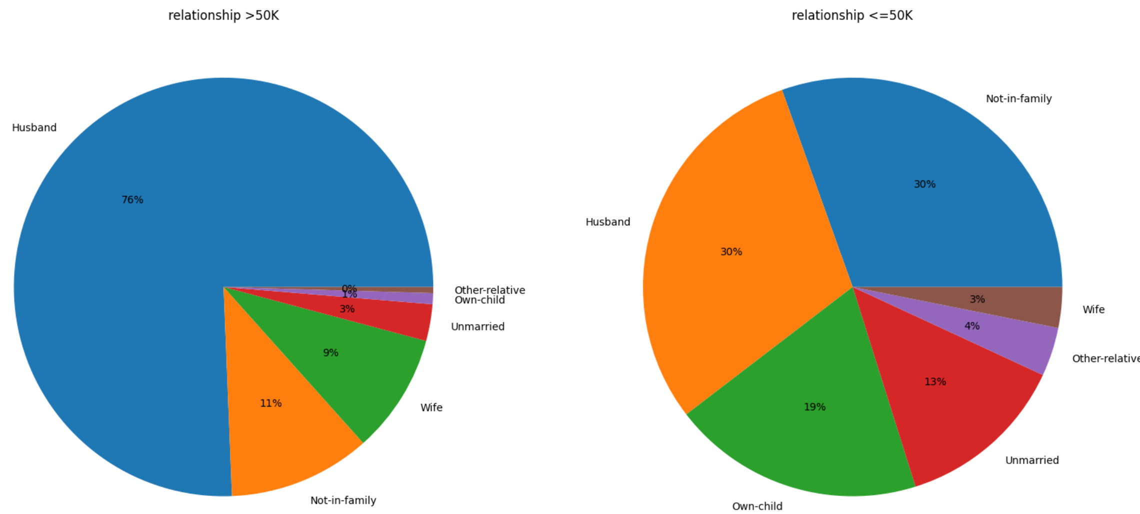


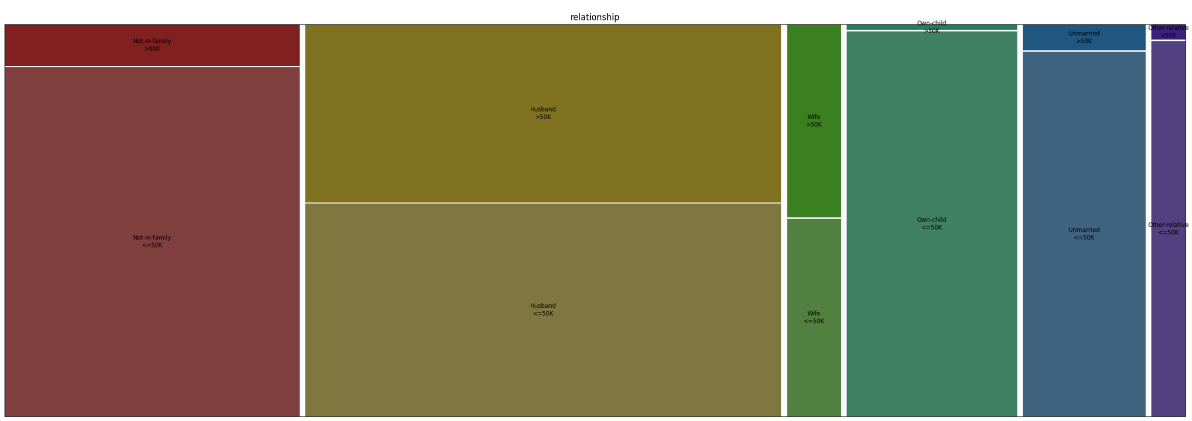





```
For variable relationship
Husband           5679
Not-in-family      823
Wife               694
Unmarried          213
Own-child           64
Other-relative      35
Name: relationship, dtype: int64
Not-in-family     6903
Husband           6784
Own-child         4402
Unmarried         2999
Other-relative     854
Wife               712
Name: relationship, dtype: int64
```

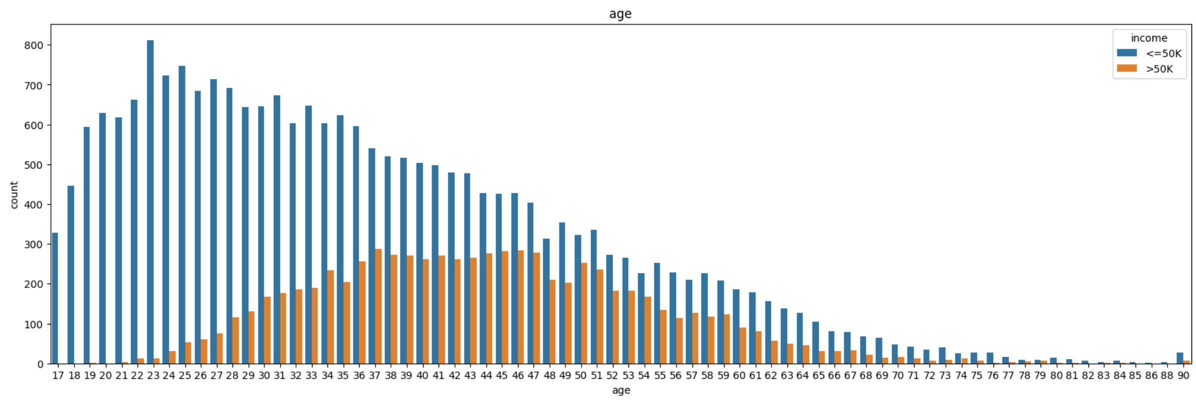


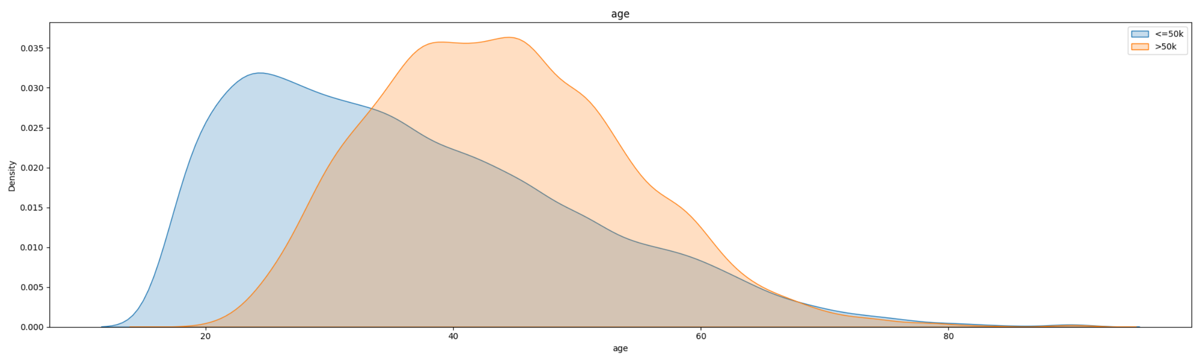


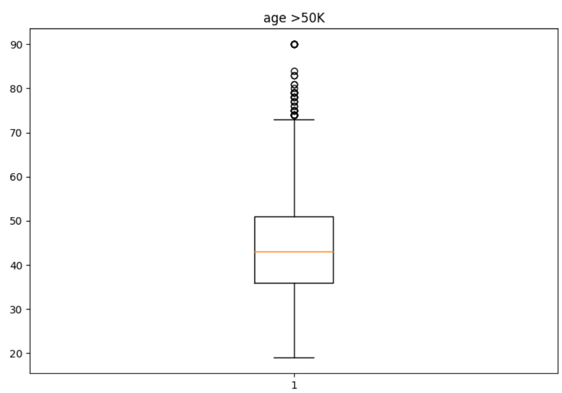


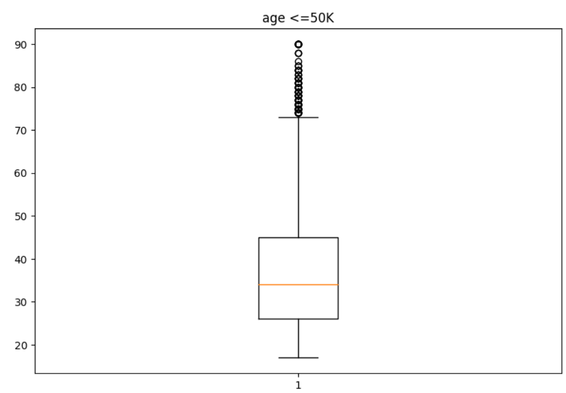
In [89]:
```
numerical('age')
numerical('education_num')
numerical('capital_gain')
numerical('capital_loss')
```

```
For variable age
Min is: 17
Max is: 90
Mean is: 38.437901995888865
Std Dev is: 13.134664776855985
```

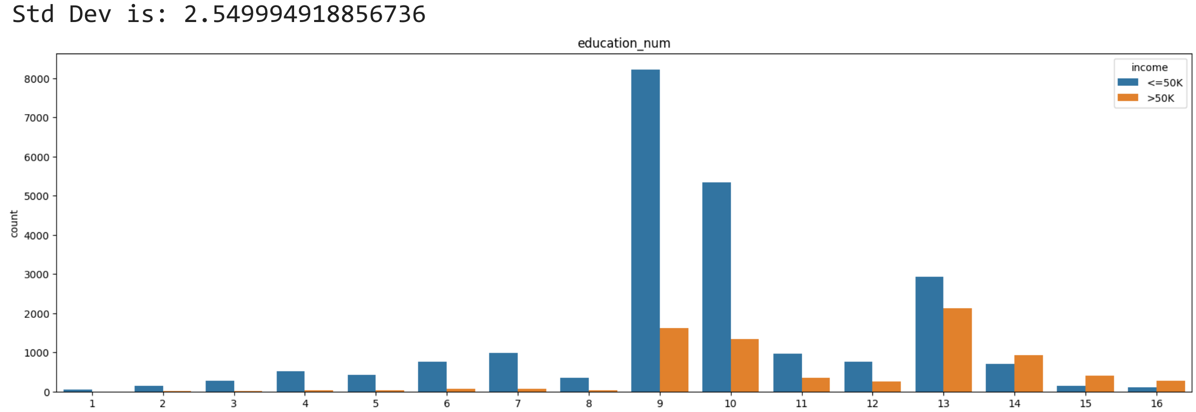


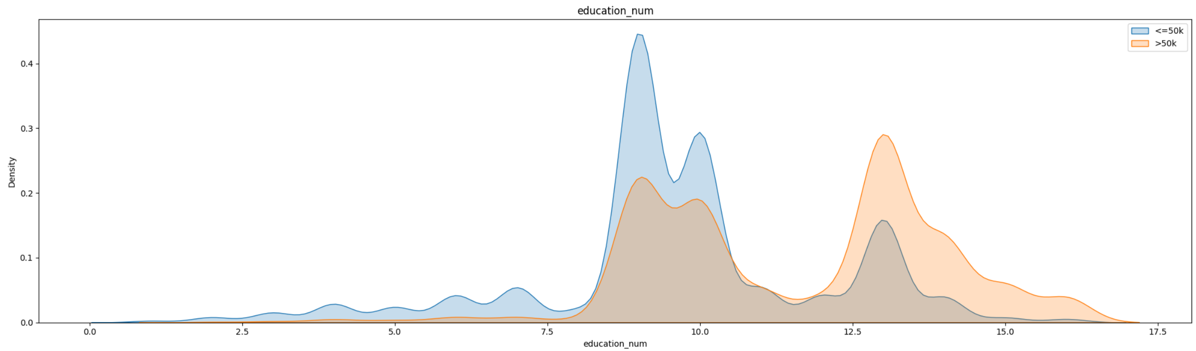


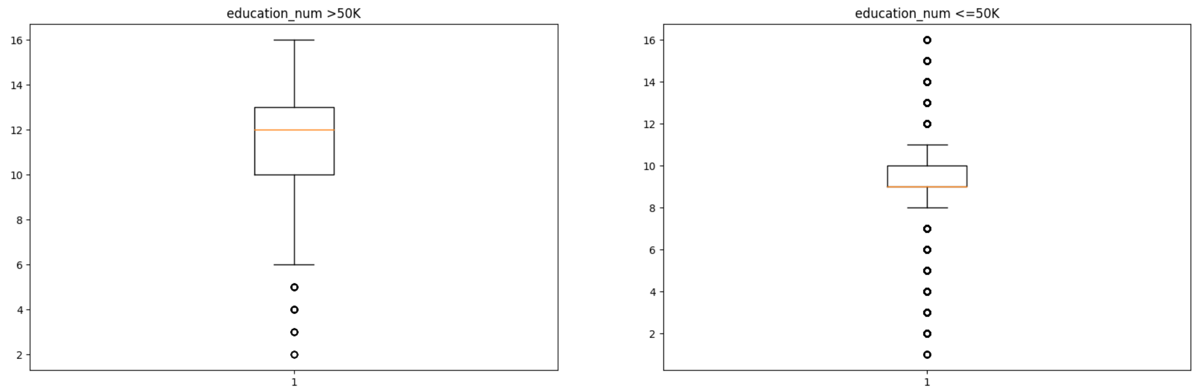




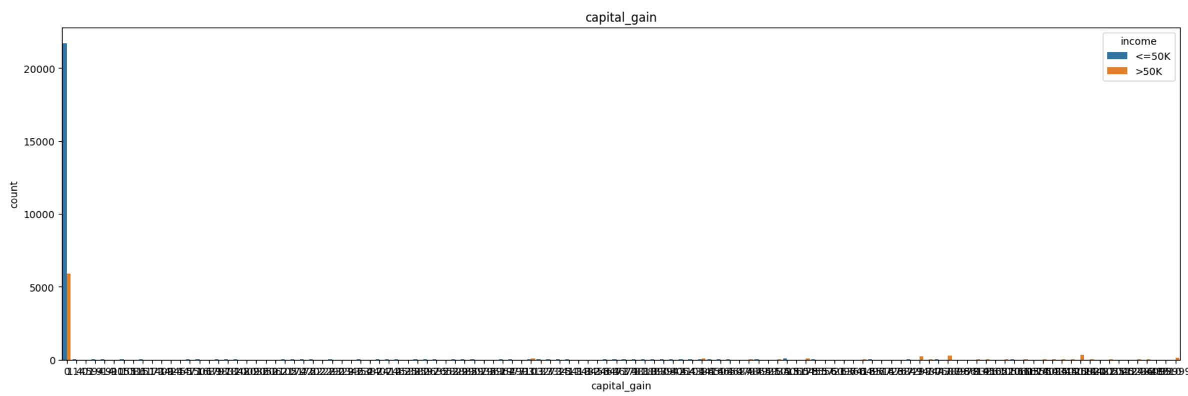

```
For variable education_num
Min is: 1
Max is: 16
Mean is: 10.12131158411246
Std Dev is: 2.549994918856736
```

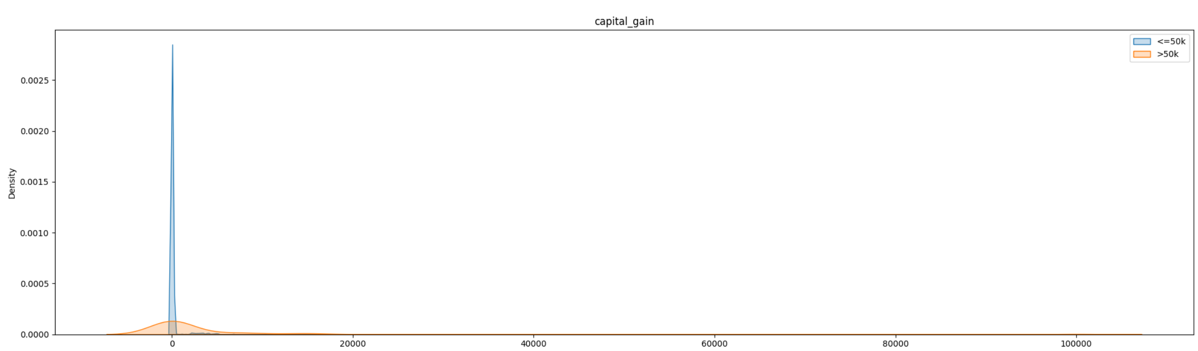


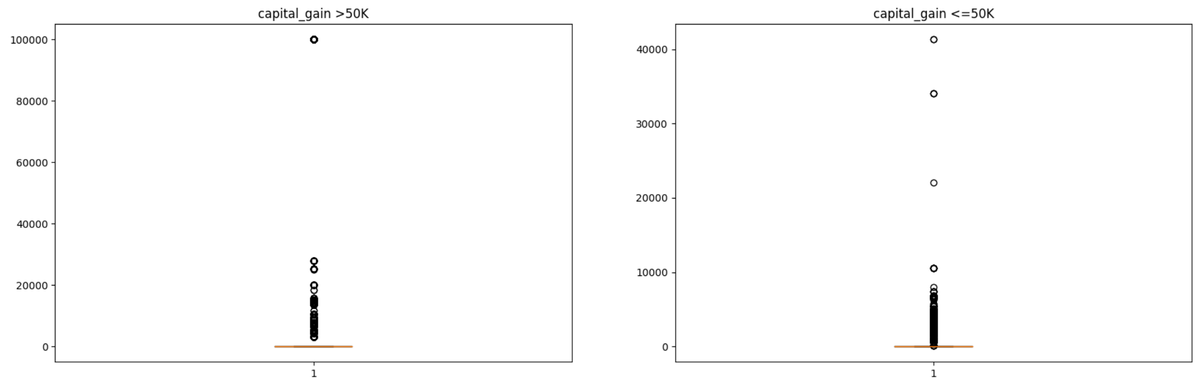




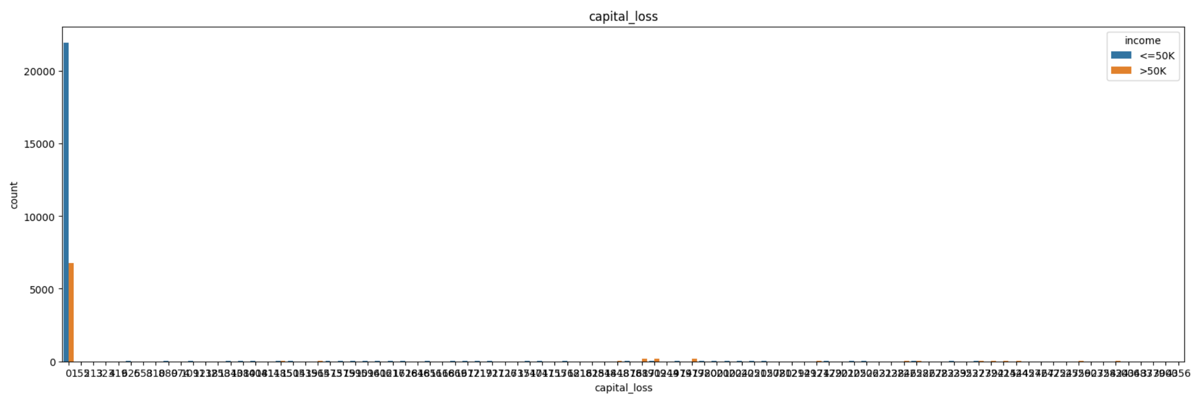

```
For variable capital_gain
Min is: 0
Max is: 99999
Mean is: 1092.0078575691268
Std Dev is: 7406.346496683503
```

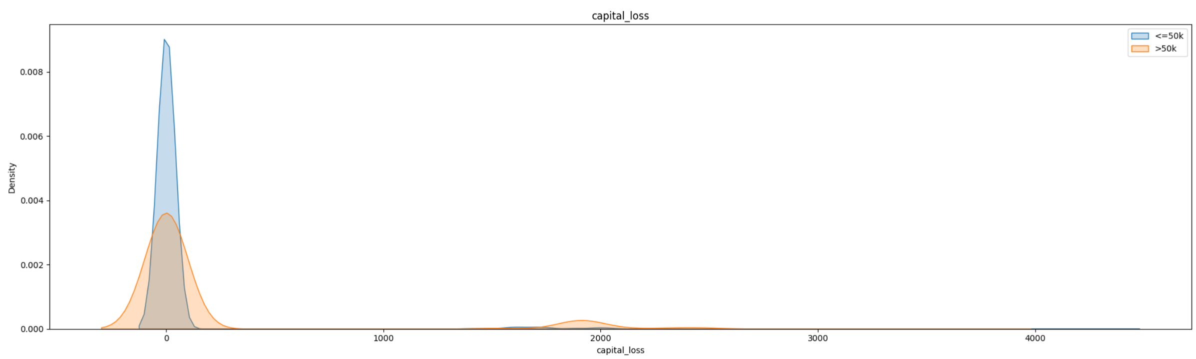


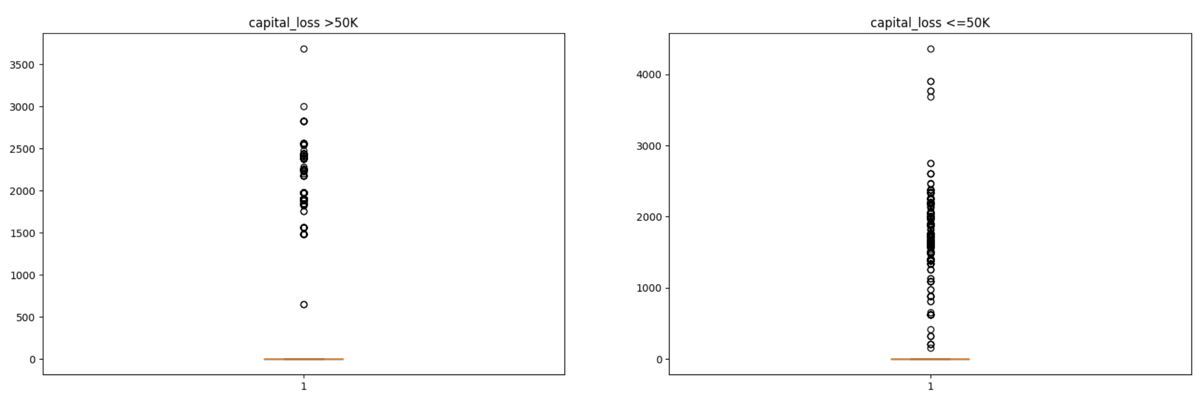





```
For variable capital_loss
Min is: 0
Max is: 4356
Mean is: 88.37248856176646
Std Dev is: 404.29837048637575
```

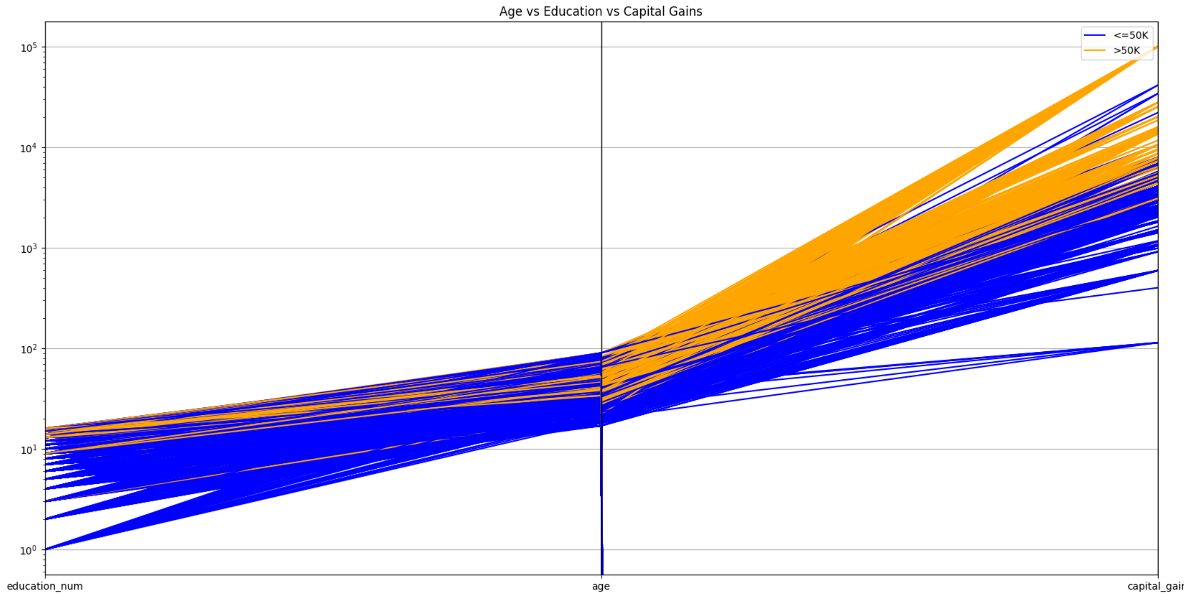






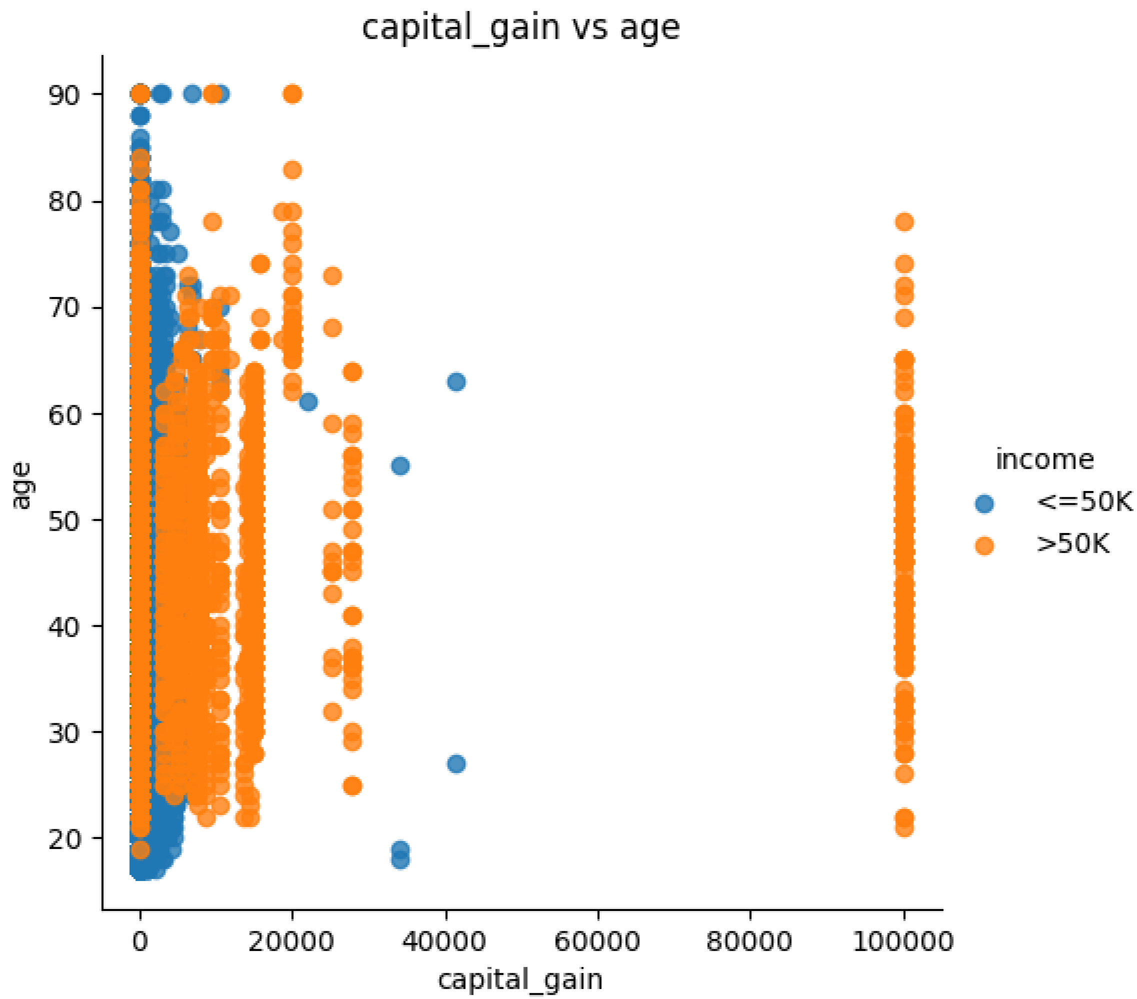
In [90]:
```
# multivariate analysis
fig, ax = plt.subplots(figsize=(20,10))
parallel_coordinates(data[[ 'education_num', 'age','capital_gain', 'income']], 'inc
plt.yscale('log')
```

```
plt.title('Age vs Education vs Capital Gains')
plt.show()
```

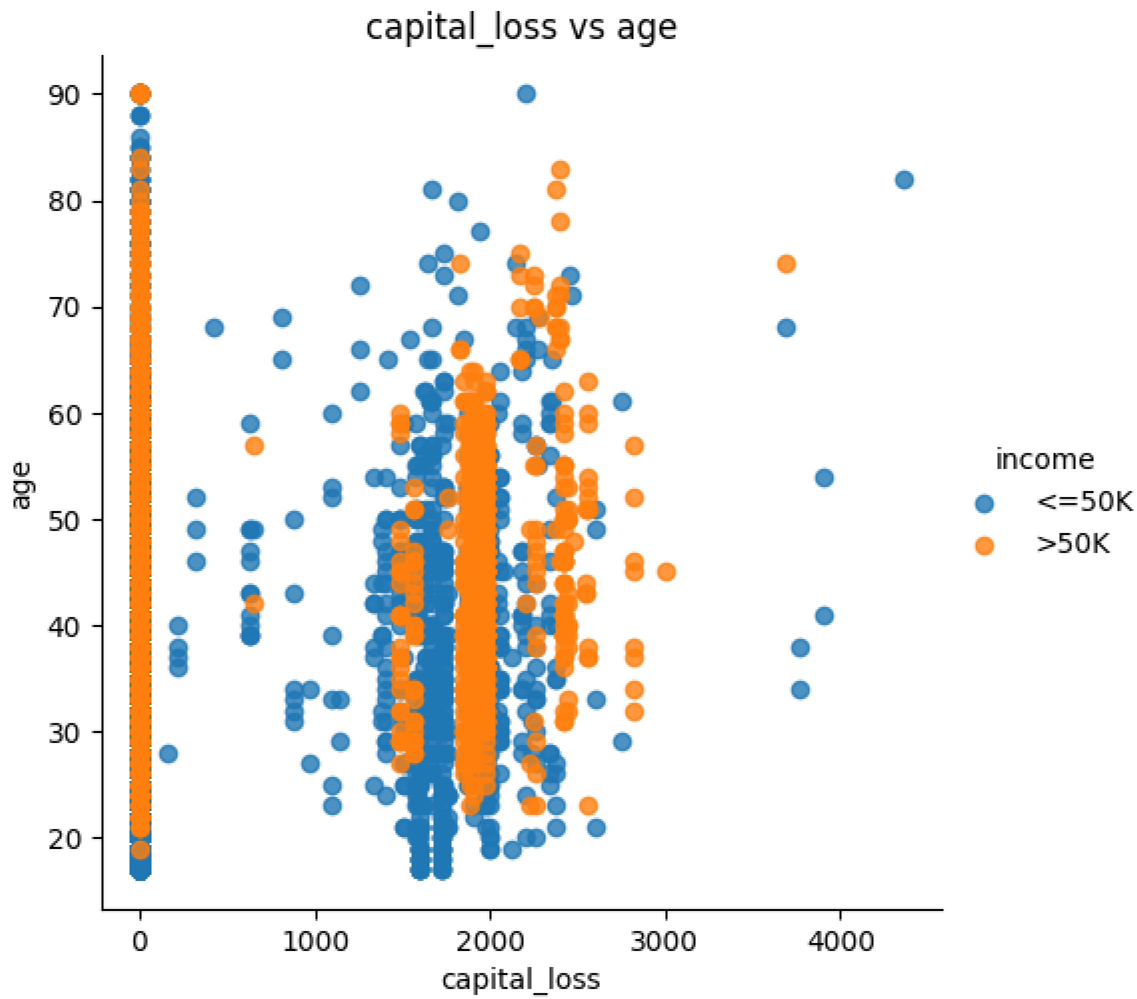


In [91]:
```
# multivariate analysis
def compare(col1, col2):
    # scatterplot of two variables by income
    fig1 = plt.figure(figsize=(20, 10))
    ax1 = sns.lmplot(x=col1, y=col2, data=data, fit_reg=False, hue='income', legend
    plt.title(col1 +' vs '+ col2)
    plt.show()

compare('capital_gain', 'age')
compare('capital_loss', 'age')
```

```
<Figure size 2000x1000 with 0 Axes>
```




```
<Figure size 2000x1000 with 0 Axes>
```


capital_loss vs age
age
90
80
70
60
50
40
30
20
0
1000
2000
3000
4000
capital_loss
income
<=50K
>50K