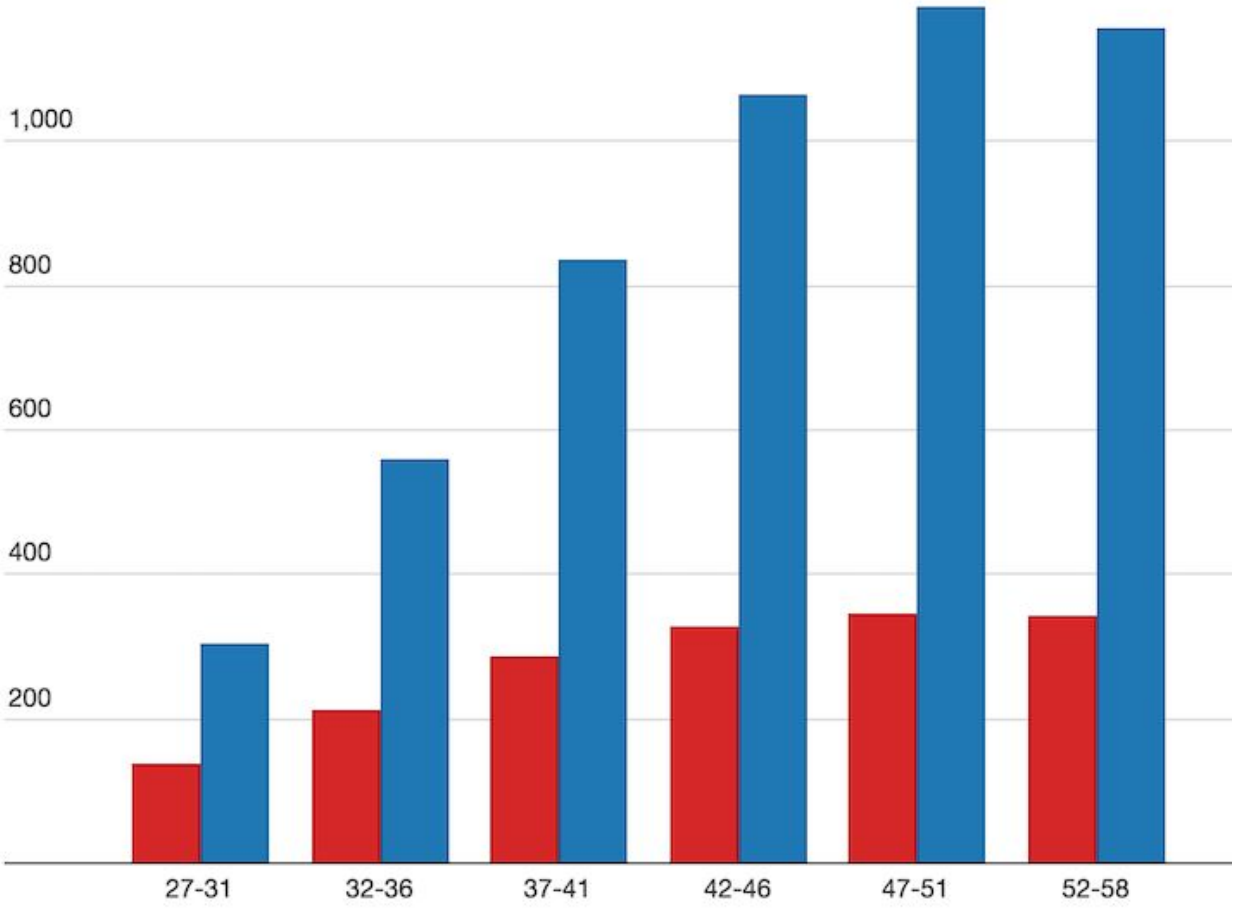
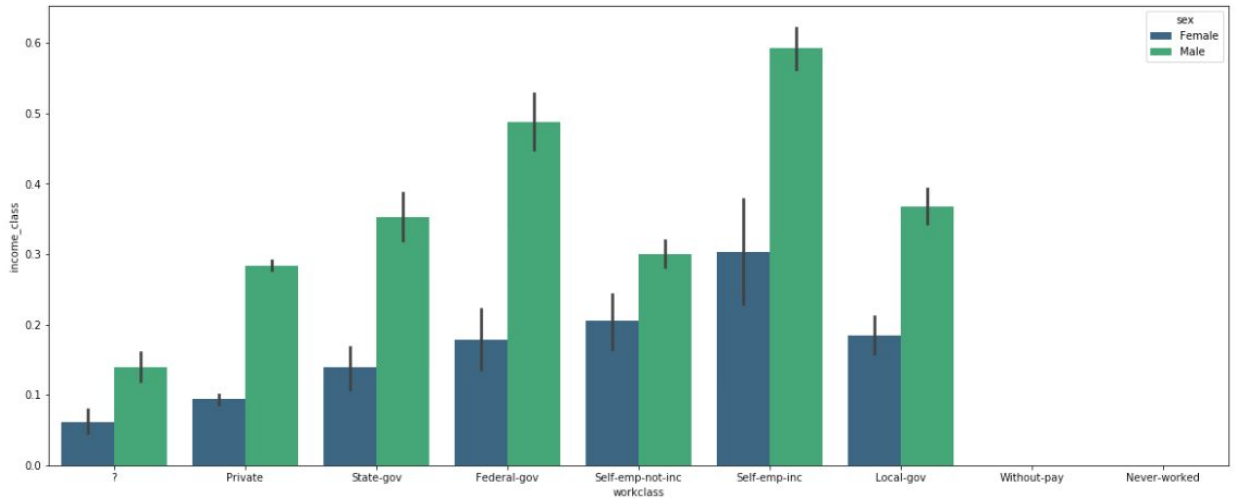
CAPSTONE2-MILESTONE 1



This report consists of data wrangling,exploratory data analysis and statistical inferences regarding adult data.

MEHMET KETENCI

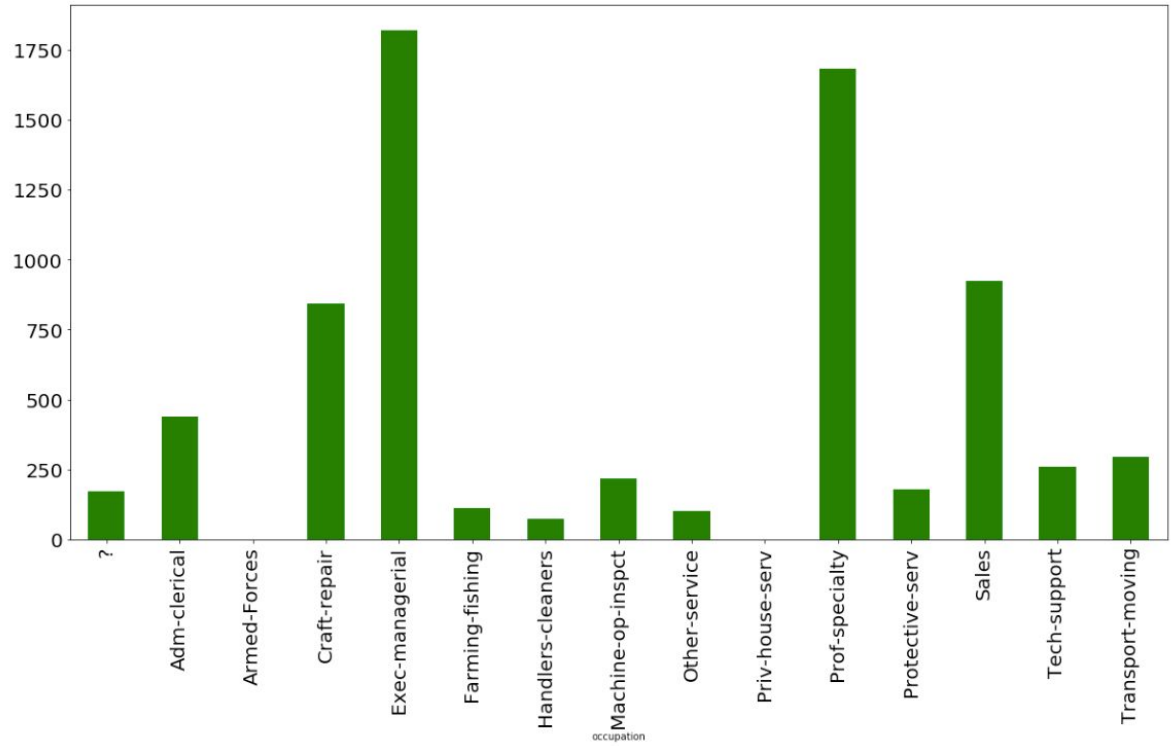
PROBLEM: Our goal is to be able to classify given data with respect to certain income level.
CLIENTS AND WHY DO THEY CARE: Result of this dataset has crucial importance for those industries below in determining tendency of group of people both high income level with low income level.
- **Marketing and e-marketing companies**:Companies may offer their different products with respect to different income levels.
- **Health, car, home and life insurance companies**: Data will be produced by statistical inferences and Eda with Machine Learning is beneficial for this industry due to mapping right customer and right insurance type and amount.
- **Investment industry**:This industry can aim right income class with respect to customers` different features.
- **Loan companies and banks**:Classified data can be used for this industry to match correct loan or credit amount with targeted people.
- **Travel agencies**: People have different vacation habits,some people may want to go to the seaside while some enjoys spending time mountains and some like museums or historical places.This industry may offer right travel options to correct people.
WHAT WILL CLIENTS DECIDE AFTER MY PROJECT: They can easily determine their target mass with simplified numbers or values and prominent and well-organized visuals.
HOW WILL I ACQUIRE THE DATA?
Data itself was obtained on http://archive.ics.uci.edu/ml/datasets/Adult



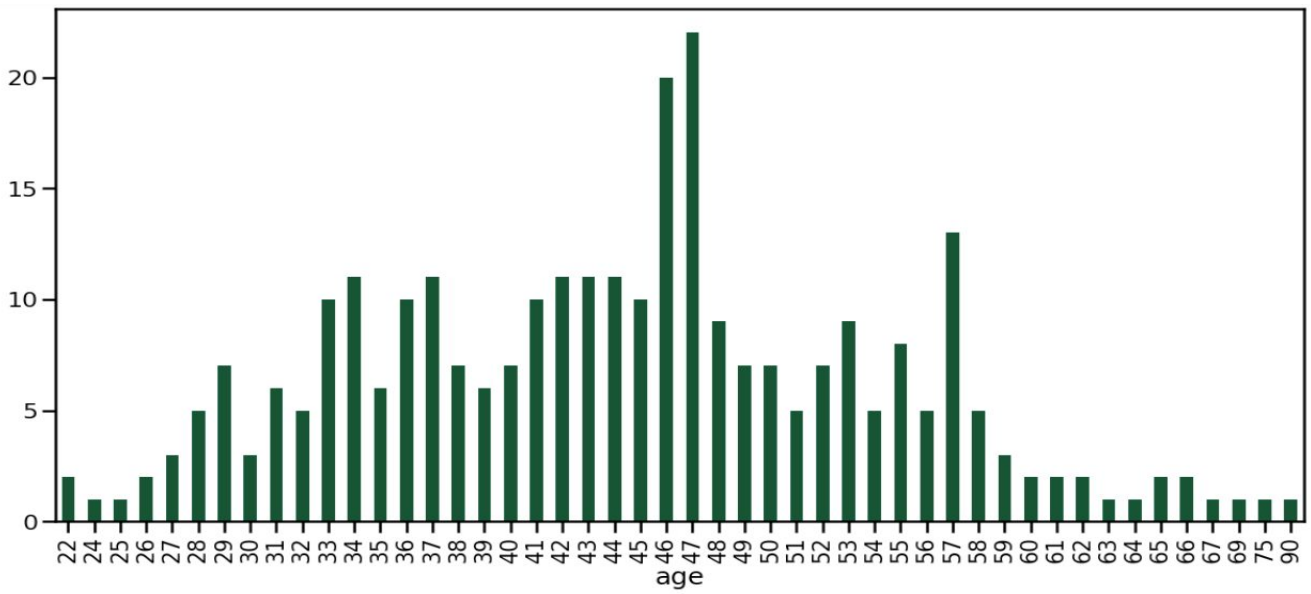
This graph shows the income distribution of groups that depend on a specific work class and gender.

```
#certain race group(white) above 50K income with respect to certain occupation
d = higher
d = d[d.race == 'White']
d.groupby(['occupation']).size().plot(kind='bar',color='green',figsize=(20,10),fontsize=20)
```

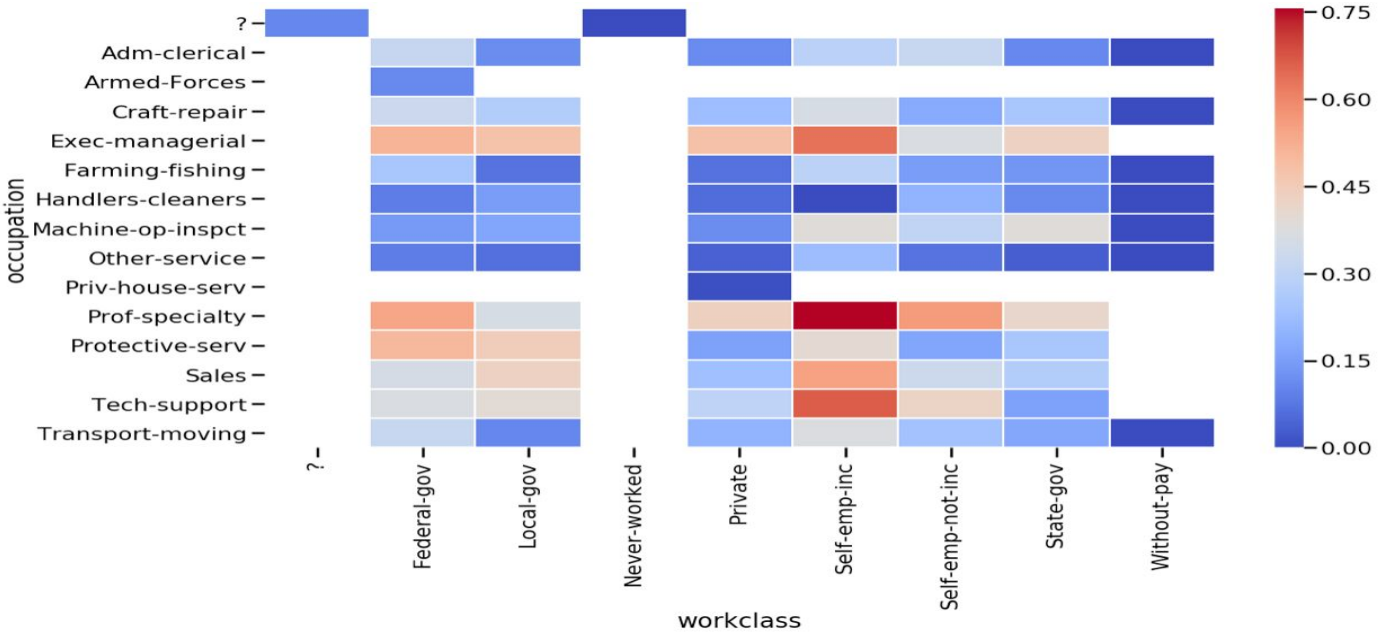
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fcdc0f2f668>
```



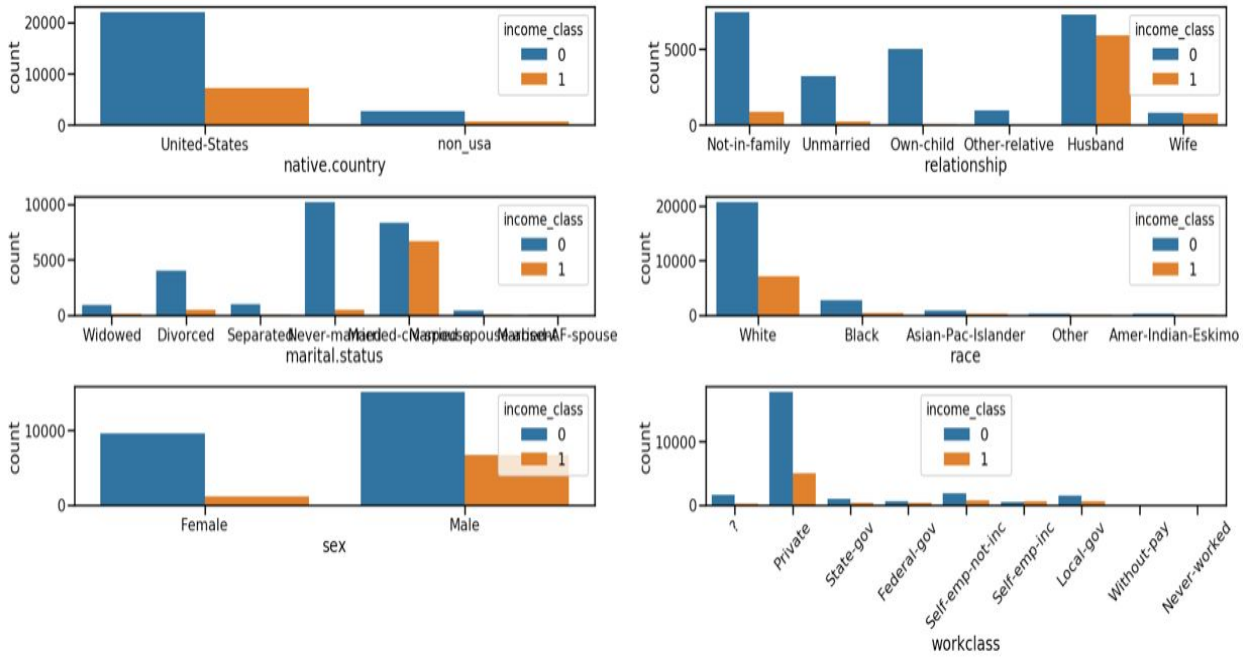
This graph shows the professional distribution of the white population with more than 50 thousand income. The top 5 consists of Exec-managerial,Prof-speciality,Sales,Craft-repair and Adm-clerical.



This graph shows the age distribution of the high-income group of black men. According to the graph, the surplus of people aged 46 and 47 is in the foreground.



Here we see the heatmap graph according to the income averages of the people working in a certain occupational group and sector and according to this graph (Prof-specialty, Self-emp-inc), (Exec-managerial, Self-emp-inc), (Tech-support, Self-emp) -inc), (Sales, Self-emp-inc) and () combinations have a high average income.



Here we also see various comparisons based on income classes.

CHI SQUARE TEST

In [ ]:
```
# H0:'Race' and 'Income' variables are independent.
# Ha:'Race' and 'Income' variables are not independent.
```

In [68]:
```
pd.crosstab(data['income'],data['race'])
```

Out[68]:

income	race	Amer-Indian-Eskimo	Asian-Pac-Islander	Black	Other	White
<=50K		275	763	2737	246	20699
>50K		36	276	387	25	7117

The null hypothesis states that knowing the race variables doesnt help us predict the variables of income. Alternative hypothesis is that knowing race variables might help us to predict income values.

In [72]:
```
from scipy.stats import chi2_contingency
chi2_contingency(pd.crosstab(data['race'],data['income']))
```

Out[72]:
```
(330.9204310085741,
 2.305960610160958e-70,
 4,
 array([[  236.10822763,    74.89177237],
        [  788.79886981,   250.20113019],
        [ 2371.71094254,   752.28905746],
        [  205.74060993,    65.25939007],
        [21117.64135008,  6698.35864992]]))
```

P-value less than 0.05(significance level), We then reject Null Hypothesis and accept alternative Hypothesis.

It means race is not an independent column.Income is correlated with race.

Importance of age for income

People make more money in time.Their income levels increase depend on accumulated years.
You will see a study supports my sentences above.

In [142]:
```
wfh = higher[higher['race']=='White']
wfh = wfh[wfh['sex']=='Female']
```

In [143]:
```
wmh = higher[higher['race']=='White']
wmh = wmh[wmh['sex']=='Male']
```

In [144]:
```
wfl = lower[lower['race']=='White']
wfl = wfl[wfl['sex']=='Female']
```

In [145]:
```
wml = lower[lower['race']=='White']
wml = wml[wml['sex']=='Male']
```

In [150]:
```
format(round(wfh['age'].mean(),2),round(wmh['age'].mean(),2),round(wfl['age'].mean(),2),round(wml['age'].mean(),2)))
```

```
average age of white females who have higher than 50K income = 42.28
average age of white males who have higher than 50K income = 44.74
average age of white females who have lower than 50K income = 36.07
average age of white males who have lower than 50K income = 37.29
```

T TEST

We are curious as to whether education level(number) differ between genders in a group that below 50K income. We take out samples from both male and female.Here is the result below:(alpha=0.05)

```
#We take samples that have below 30 for the sake of being independent rule
#from both group in order to apply T test.
a = np.random.choice(lf['education.num'],24)
b = np.random.choice(lm['education.num'],20)
print(a.mean(),b.mean(),a.std(),b.std())
```

```
9.875 9.95 2.8182810955143087 2.108909670896314
```

```
pd.DataFrame({'Mean':[10.5,9.45],
              'Std':[2.33,2.65],
             'Samples':[24,20]},
             index=['Female<50K','Male<50K'])
```

	Mean	Std	Samples
Female<50K	10.50	2.33	24
Male<50K	9.45	2.65	20

```
t=(10.50-9.45)/((2.33**2/24)+(2.65**2/20))**0.5
t


#Small t value indicates that education levels for male and female are similar.
```

```
1.3819029685511863
```

```
stats.ttest_ind(a,b)
```

```
Ttest_indResult(statistic=-0.09601321720529037, pvalue=0.9239667474636487)
```

We fail to reject Null hypothesis since p value is greater than threshold(significance level). As a result,there is no significance difference between being opposite genders of lower income as to education levels.

Z TEST

2) Comparing average age of two different groups

Average age of black people with above 50K and below with 50K in the dataset are respectively 43 and 36. standard deviations of both group are respectively 9 and 12. Test if mean of both group in real world is equal each other?(alpha=0.01)

```
#   H0: Ma-Mb = 0
#   Ha: Ma-Mb ≠ 0
```

```
Ma = 43
Mb = 36
σ1 = 9
σ2 = 12
z = (Ma-Mb-0)/((σ1**2/387)+(σ2**2/2737))**0.5
z
```

```
13.677859045246745
```

```
our z value=13.7 is way bigger than z critical value=2.575
so we reject Null hypothesis and accept alternative hypothesis
As a result mean age of black group with above 50K and below 50K are not going to be equal e
ach other.
```