P1_Hromada_Alex

January 31, 2021

1 P1

1.1 Alex Hromada

[50]:
```
import pandas as pd
import numpy as np
import matplotlib as mpl
import matplotlib.pyplot as plt
import seaborn as sb
%matplotlib inline
```

[118]:
```
URL = "http://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.
 ↪data"
URL2 = "D:/School/spring-2021/cs4821/p1/P1-data/nfl-20-running-stats.csv"

#read dataset
adult = pd.read_csv(URL, header=None, names=['age', 'workclass', 'fnlwgt',␣
 ↪'education',
                                              'education-num', 'marital-status',␣
 ↪'occupation',
                                             'relationship', 'race', 'sex',␣
 ↪'capital-gain',
                                             'capital-loss', 'hours-per-week',␣
 ↪'native-country',
                                             'income'])

nfl = pd.read_csv(URL2, header=None, names=['Rk', 'Player', 'Tm', 'Age',
                                             'Pos', 'G', 'GS', 'Att',
                                             'Yds', 'TD', '1D', 'Lng',
                                             'Y/A', 'Y/G', "Fmb"], skiprows=[0])


nfl.head()
```

[118]:

	Rk	Player	Tm	Age	Pos	G	GS	Att	Yds	TD	1D	\
0	1	Derrick Henry *\HenrDe00	TEN	26	RB	16	16	378	2027	17	98	
1	2	Dalvin Cook*\CookDa01	MIN	25	RB	14	14	312	1557	16	91	
2	3	Josh Jacobs*\JacoJo01	LVR	22	RB	15	15	273	1065	12	61	
3	4	David Montgomery\MontDa01	CHI	23	RB	15	14	247	1070	8	59	

```
4   5   Ezekiel Elliott\ElliEz00  DAL   25  RB  15  15  244   979   6  62

   Lng  Y/A    Y/G  Fmb
0   94  5.4  126.7    3
1   70  5.0  111.2    5
2   28  3.9   71.0    2
3   80  4.3   71.3    1
4   31  4.0   65.3    6
```

2 Q1

2.1 Q1(a)

AGE is the age of an individual as reported by that person for the 1990 census; the value is reported in integer units of years.

WORKCLASS is the type of work/employment of an individual as reported by that person for the 1990 census; the value is reported as a string descripter.

FNLWGT is a weight value of an individual based on specified socio-economic characteristics that is prepared by the Census Bureau; the value is reported as an integer value.

EDUCATION is the level of education of an individual as reported by that person for the 1990 census; the value is reported as a string descriptor.

EDUCATION-NUM is the number of years of education of an individual as reported by that person for the 1990 census; the value is reported in integer units of years.

MARITAL-STATUS is the marital status of an individual as reported by that person for the 1990 census; the value is reported as a string descriptor.

OCCUPATION is the occupation of an individual as reported by that person for the 1990 census; the value is reported as a string descriptor.

RELATIONSHIP is the relationship of an individual relative to others as reported by that person for the 1990 census; the value is reported as a string descriptor

RACE is the race of an individual as reported by that person for the 1990 census; the value is reported as a string descriptor.

SEX is the sex of an individual as reported by that person for the 1990 census; the value is reported as a string descriptor.

CAPITAL-GAIN is the amount of capital gain of an individual as reported by that person for the 1990 census; the value is reported in integer units of dollars.

CAPITAL-LOSS is the amount of capital loss of an individual as reported by that person for the 1990 census; the value is reported in integer units of dollars.

HOURS-PER-WEEK is the average hours worked per week of an individual as reported by that person for the 1990 census; the value is reported in integer units of hours.

NATIVE-COUNTRY is the country of origin of an individual as reported by that person for the 1990 census; the value is reported as a string descriptor.

2.2 Q1(b)

2.2.1 Q1(b)(i)

Missing data in the dataset is represented by a '?'.

2.2.2 Q1(b)(ii)

[62]:
```
# adult_missing = adult.isin(['?'])



# adult_missing.loc[[27]]



# adult_missing = adult.isna()

# adult_missing = adult_missing.sum()
# adult_missing
```

[62]:
```
      age  workclass  fnlwgt  education  education-num  marital-status  \
27  False      False   False      False          False           False

    occupation  relationship   race    sex  capital-gain  capital-loss  \
27       False         False  False  False         False         False

    hours-per-week  native-country  income
27           False           False   False
```

[70]:
```
adult.loc[[27]]
```

[70]:
```
    age workclass  fnlwgt      education  education-num       marital-status  \
27   54         ?  180211   Some-college             10   Married-civ-spouse

   occupation relationship                 race    sex  capital-gain  \
27          ?      Husband   Asian-Pac-Islander   Male             0

    capital-loss  hours-per-week native-country income
27             0              60          South   >50K
```

2.3 Q1(c)

Variables of a numeric datatype in this dataset include AGE, FNLWGT, EDUCATION-NUM, CAPITAL-GAIN, CAPITAL-LOSS, HOURS-PER-WEEK.

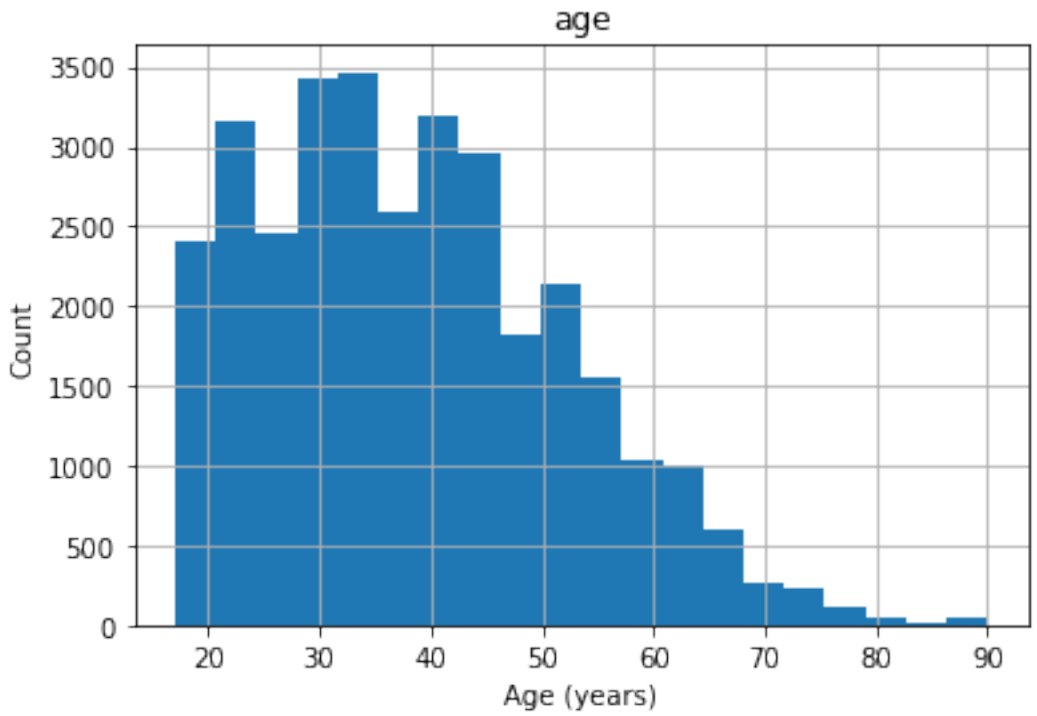
Variables of a categorial datatype in this dataset include WORKCLASS, EDUCATION, MARITAL-STATUS, OCCUPATION, RELATIONSHIP, RACE, SEX, NATIVE-COUNTRY.

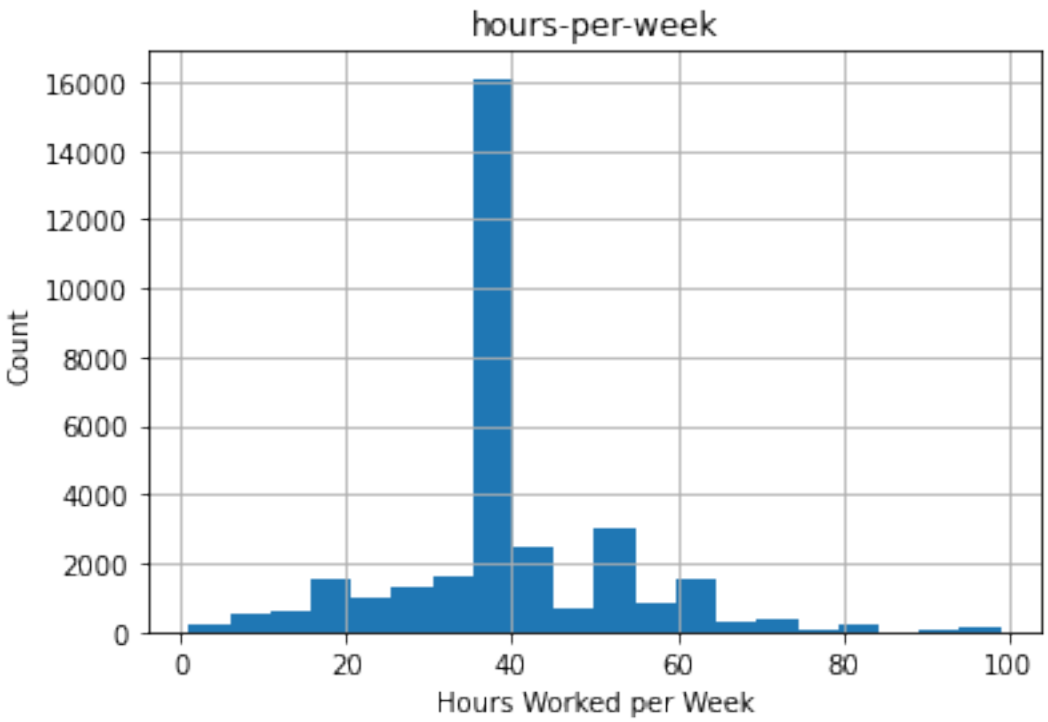
2.4 Q1(d)

2.4.1 Q1(d)(i)

[153]:
```
adult.hist(column='age', bins=20)
plt.xlabel("Age (years)")
plt.ylabel("Count");

adult.hist(column='hours-per-week', bins=20)
plt.xlabel("Hours Worked per Week")
plt.ylabel("Count");
```



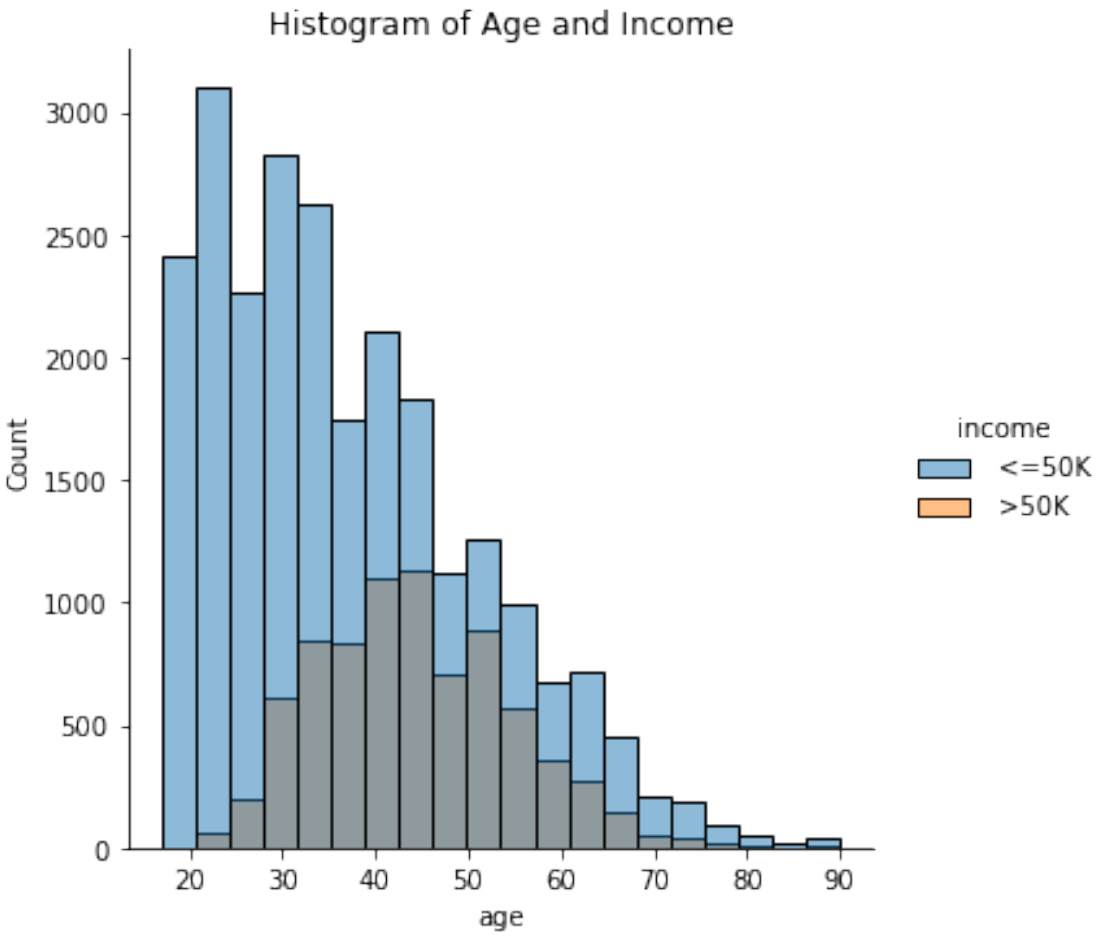


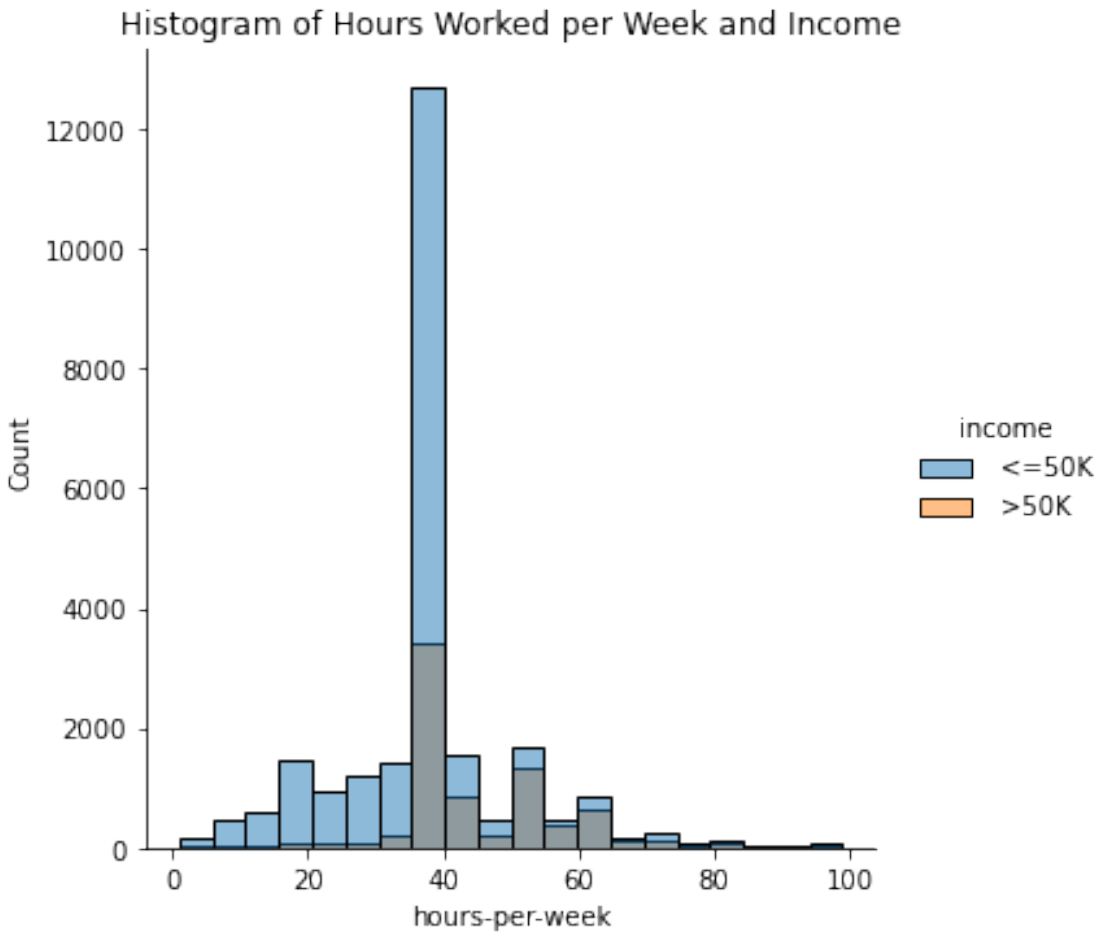
2.4.2 Q1(d)(ii)

[83]:
```
sb.displot(adult, x="age", hue="income", bins=20)
plt.title("Histogram of Age and Income")

sb.displot(adult, x="hours-per-week", hue="income", bins=20)
plt.title("Histogram of Hours Worked per Week and Income")
```

[83]:
```
Text(0.5, 1.0, 'Histogram of Hours Worked per Week and Income')
```

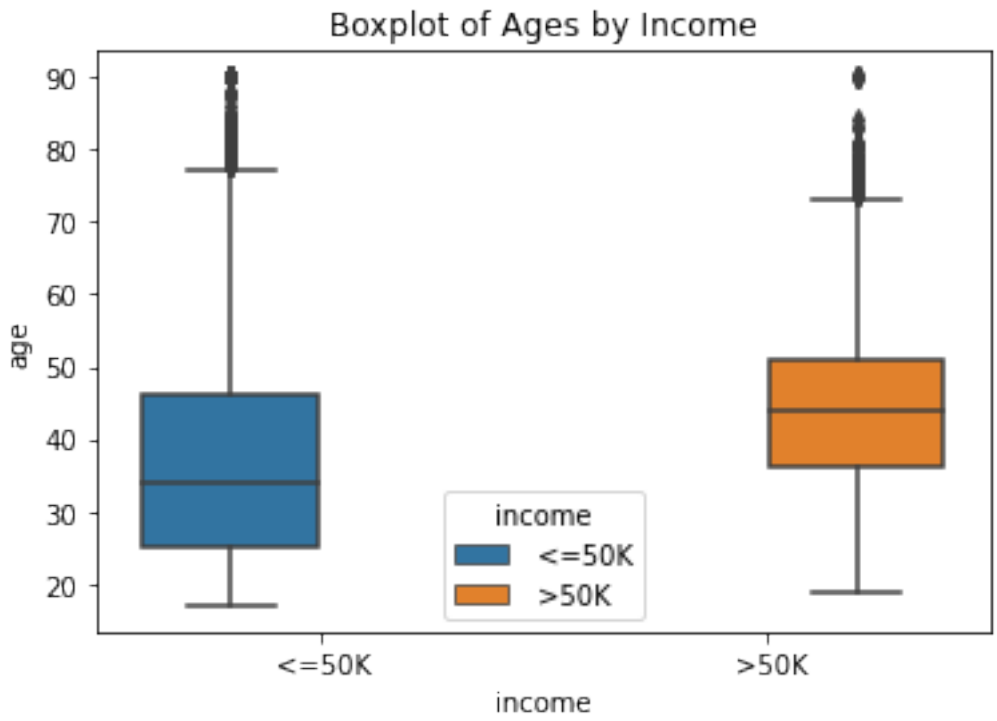

Histogram of Age and Income
Count
3000
2500
2000
1500
1000
500
0
20
30
40
50
60
70
80
90
age
income
<=50K
>50K



2.4.3 Q1(d)(iii)

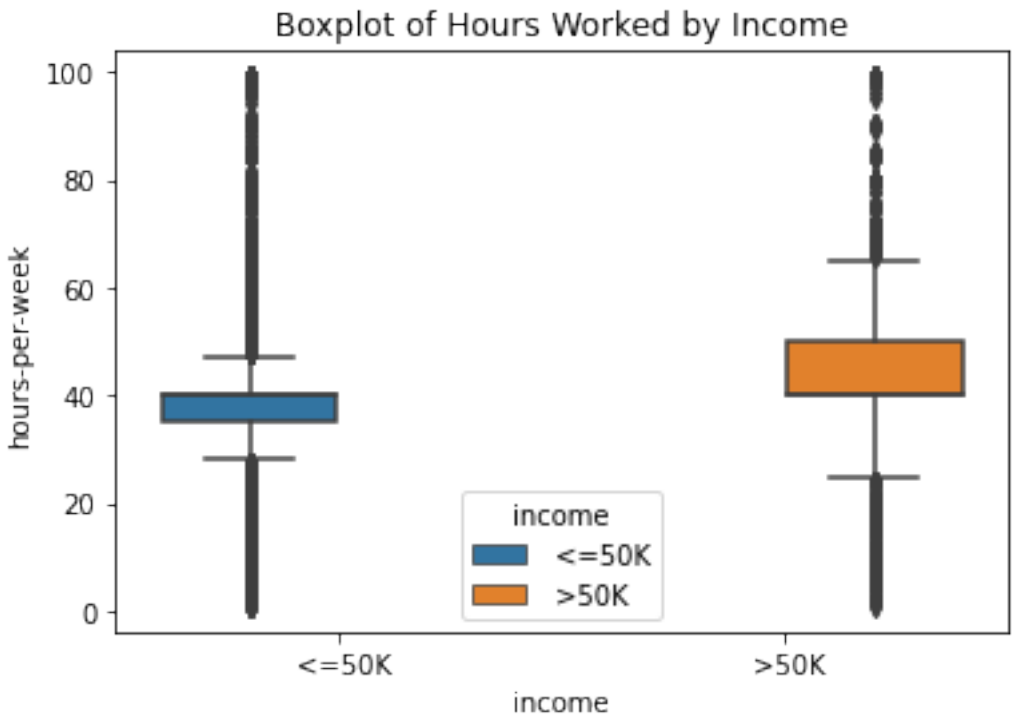
[90]:
```
sb.boxplot(x='income',
           y='age',
           hue='income',
           data=adult[['age', 'income']])
plt.title("Boxplot of Ages by Income")
```

[90]: Text(0.5, 1.0, 'Boxplot of Ages by Income')



[89]:
```
sb.boxplot(x='income',
           y='hours-per-week',
           hue='income',
           data=adult[['hours-per-week', 'income']])
plt.title("Boxplot of Hours Worked by Income")
```

[89]: Text(0.5, 1.0, 'Boxplot of Hours Worked by Income')



2.4.4 Q1(d)(iv)

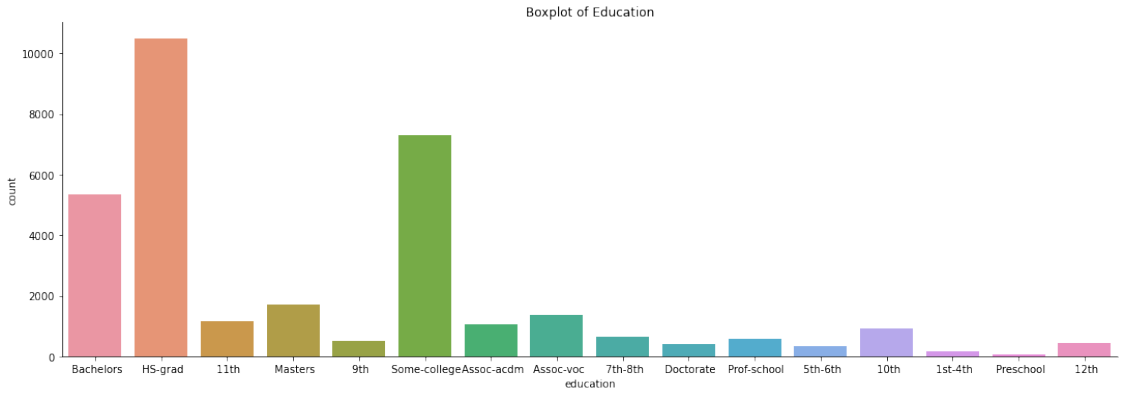
Plotting the data of ages with income groups of *>50k* and *<=50k* reveals a right skewed distribution with a large count of individuals under the age of 45 making less than \$50k a year, and a lesser number of the older age groups making an income of less than \$50k a year. It is also revealed that the group of individuals making an income of less than \$50k has a median age of around 35, compared to the group of individuals making \$50k or more which has a median age of around 45. There is an overall lower count of individuals making an income of \$50k or more. The median hours worked for both groups is centered arounf 40, however, the group making \$50k or more has a higher interquartile range than the group making less than \$50k. Both groups have a large number of outliers due to the heavy amount of individuals with 40 hour work weeks.

2.5 Q1(e)

2.5.1 Q1(e)(i)

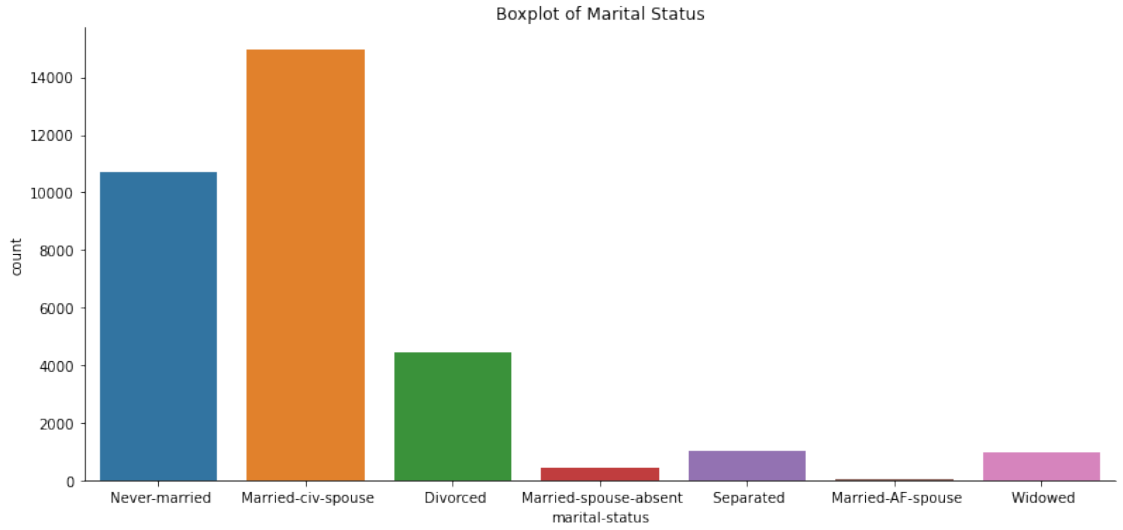
[154]:
```
sb.catplot(x="education", kind="count", aspect=3, data=adult)
plt.title("Boxplot of Education")
```

[154]:
```
Text(0.5, 1.0, 'Boxplot of Education')
```



[155]:
```
sb.catplot(x="marital-status", kind="count", aspect=2.2, data=adult)
plt.title("Boxplot of Marital Status")
```

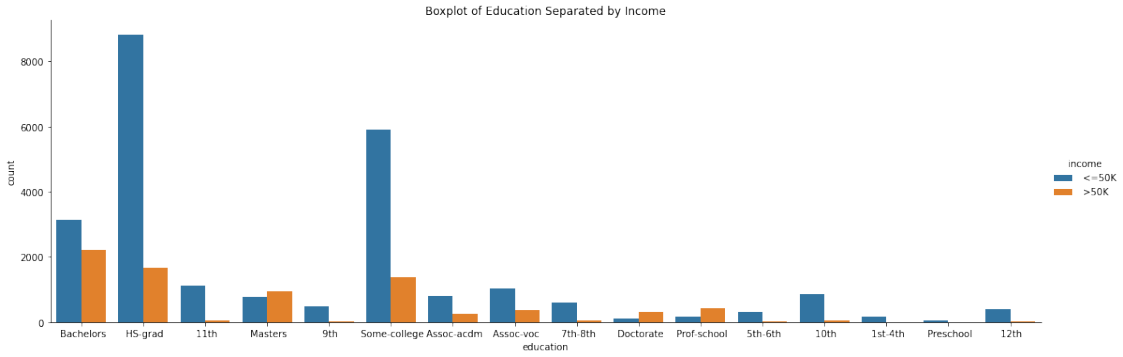
[155]: Text(0.5, 1.0, 'Boxplot of Marital Status')



2.5.2 Q1(e)(ii)

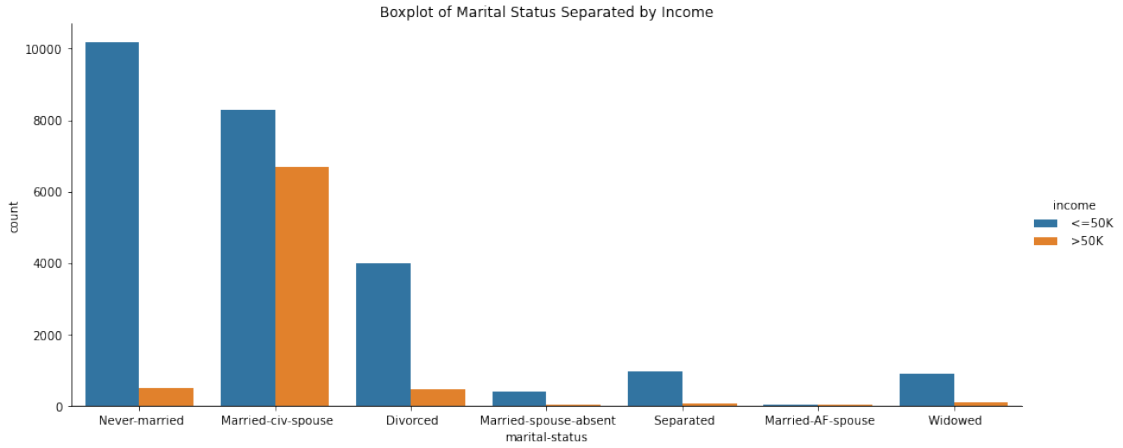
[111]:
```
sb.catplot(x="education", hue="income", kind="count", aspect=3.1, data=adult)
plt.title("Boxplot of Education Separated by Income")
```

[111]: Text(0.5, 1.0, 'Boxplot of Education Separated by Income')



[114]:
```
sb.catplot(x="marital-status", hue="income", kind="count", aspect=2.4,␣
 ↪data=adult)
plt.title("Boxplot of Marital Status Separated by Income")
```

[114]:
```
Text(0.5, 1.0, 'Boxplot of Marital Status Separated by Income')
```



2.5.3 Q1(e)(iii)

Plotting the data of highest level of education reveals that the majority of individuals of this dataset have only a high school education, some college education, or a bachelor's degree. The individuals with only a high school education or some college education are primarily in the group making less than \$50k in income, as well as the majority of individuals in this dataset who have a bachelor's degree. Individuals in this dataset are primarily married with a civilian spouse or never married, with the individuals reported as never married primarily making less than \$50k in income, where most indivduals in the group making \$50k or more in income are married with a civilian spouse. However, the majority of individuals in this dataset that are married with a civilian spouse make under \$50k in income.

3 Q2

[147]:
```
conditions = [
    (nfl['TD'] < 5),
    (nfl['TD'] >= 5)
]

values = ['<5', '>=5']

nfl['TD/5'] = np.select(conditions, values)

nfl
```

[147]:
```
      Rk                       Player   Tm  Age  Pos   G  GS  Att   Yds  TD  \
0      1    Derrick Henry *\HenrDe00  TEN   26   RB  16  16  378  2027  17
1      2       Dalvin Cook*\CookDa01  MIN   25   RB  14  14  312  1557  16
2      3       Josh Jacobs*\JacoJo01  LVR   22   RB  15  15  273  1065  12
3      4   David Montgomery\MontDa01  CHI   23   RB  15  14  247  1070   8
4      5    Ezekiel Elliott\ElliEz00  DAL   25   RB  15  15  244   979   6
..   ...                         ...  ...  ...  ...  ..  ..  ...   ...  ..
367  368  Jonathan Williams\WillJo07  DET   26  NaN   5   0    1     5   0
368  369      Mike Williams\WillMi07  LAC   26   WR  15  11    1     1   0
369  370         Javon Wims\WimsJa00  CHI   26   wr  13   1    1     2   0
370  371  Olamide Zaccheaus\ZaccOl01  ATL   23   wr  11   2    1     0   0
371  372    Brandon Zylstra\ZylsBr00  CAR   27   wr  16   2    1     1   0

     1D  Lng  Y/A    Y/G  Fmb TD/5
0    98   94  5.4  126.7    3  >=5
1    91   70  5.0  111.2    5  >=5
2    61   28  3.9   71.0    2  >=5
3    59   80  4.3   71.3    1  >=5
4    62   31  4.0   65.3    6  >=5
..   ..  ...  ...    ...  ...  ...
367   0    5  5.0    1.0    1   <5
368   0    1  1.0    0.1    0   <5
369   0    2  2.0    0.2    0   <5
370   0    0  0.0    0.0    0   <5
371   0    1  1.0    0.1    0   <5

[372 rows x 16 columns]
```

[148]:
```
rb6 = nfl[((nfl["Pos"] == "RB") | (nfl["Pos"] == "FB")) & (nfl["G"] >= 6)]

rb6.head()
```

[148]:
```
   Rk                     Player   Tm  Age Pos   G  GS  Att   Yds  TD  1D  \
0   1  Derrick Henry *\HenrDe00  TEN   26  RB  16  16  378  2027  17  98
```

```
1   2      Dalvin Cook*\CookDa01  MIN   25  RB  14  14  312  1557  16  91
2   3      Josh Jacobs*\JacoJo01  LVR   22  RB  15  15  273  1065  12  61
3   4  David Montgomery\MontDa01  CHI   23  RB  15  14  247  1070   8  59
4   5   Ezekiel Elliott\ElliEz00  DAL   25  RB  15  15  244   979   6  62

   Lng  Y/A    Y/G  Fmb TD/5
0   94  5.4  126.7    3  >=5
1   70  5.0  111.2    5  >=5
2   28  3.9   71.0    2  >=5
3   80  4.3   71.3    1  >=5
4   31  4.0   65.3    6  >=5
```

3.1 Q2(a)

[130]:
```
nflRB6NumSample = rb6['Player'].count()
print("Sample size of players that are running back or full back with 6 or more␣
 ↪games:", nflRB6NumSample)
```

```
Sample size of players that are running back or full back with 6 or more games:
32
```

3.2 Q2(b)

[136]:
```
nflTDMean = rb6['TD'].mean()
nflTDMedian = rb6['TD'].median()
nflTDMode = rb6['TD'].mode()
nflFmbMean = rb6['Fmb'].mean()
nflFmbMedian = rb6['Fmb'].median()
nflFmbMode = rb6['Fmb'].mode()

print("TD Mean: %.3f, TD Median; %.3f, TD Mode: %.3f" % (nflTDMean,␣
 ↪nflTDMedian, nflTDMode))
print("Fmb Mean: %.3f, Fmb Median: %.3f, Fmb Mode: %.3f" % (nflFmbMean,␣
 ↪nflFmbMedian, nflFmbMode))
```

```
TD Mean: 6.969, TD Median; 6.000, TD Mode: 6.000
Fmb Mean: 1.750, Fmb Median; 1.000, Fmb Mode: 1.000
```

3.3 Q2(c)

[142]:
```
nflTD1stQuart = rb6['TD'].quantile(0.25)
nflTD3rdQuart = rb6['TD'].quantile(0.75)
nflTD37thPercent = rb6['TD'].quantile(0.37)
nflFmb1stQuart = rb6['Fmb'].quantile(0.25)
nflFmb3rdQuart = rb6['Fmb'].quantile(0.75)
nflFmb37thPercent = rb6['Fmb'].quantile(0.37)
```

```
print("TD First Quartile: %.3f, TD Third Quartile: %.3f, TD 37th Percentile: %.
↪3f" % (nflTD1stQuart, nflTD3rdQuart, nflTD37thPercent) )
print("Fmb First Quartile: %.3f, Fmb Third Quartile: %.3f, Fmb 37th Percentile:␣
↪%.3f" % (nflFmb1stQuart, nflFmb3rdQuart, nflFmb37thPercent) )
```

```
TD First Quartile: 3.750, TD Third Quartile: 9.250, TD 37th Percentile: 6.000
Fmb First Quartile: 1.000, Fmb Third Quartile: 2.000, Fmb 37th Percentile: 1.000
```

3.4 Q2(d)

[143]:
```
nflYGMin = rb6['Y/G'].min()
nflYG1stQuart = rb6['Y/G'].quantile(0.25)
nflYGMedian = rb6['Y/G'].median()
nflYG3rdQuart = rb6['Y/G'].quantile(0.75)
nflYGMax = rb6['Y/G'].max()

nflLngMin = rb6['Lng'].min()
nflLng1stQuart = rb6['Lng'].quantile(0.25)
nflLngMedian = rb6['Lng'].median()
nflLng3rdQuart = rb6['Lng'].quantile(0.75)
nflLngMax = rb6['Lng'].max()

summary = {
    '': ['Y/G', 'Lng'],
    'Min': [nflYGMin, nflLngMin],
    "1st Quartile": [nflYG1stQuart, nflLng1stQuart],
    "Median": [nflYGMedian, nflLngMedian],
    "3rd Quartile": [nflYG3rdQuart, nflLng3rdQuart],
    "Max": [nflYGMax, nflLngMax]
}

sum5num = pd.DataFrame(data=summary)
sum5num
```

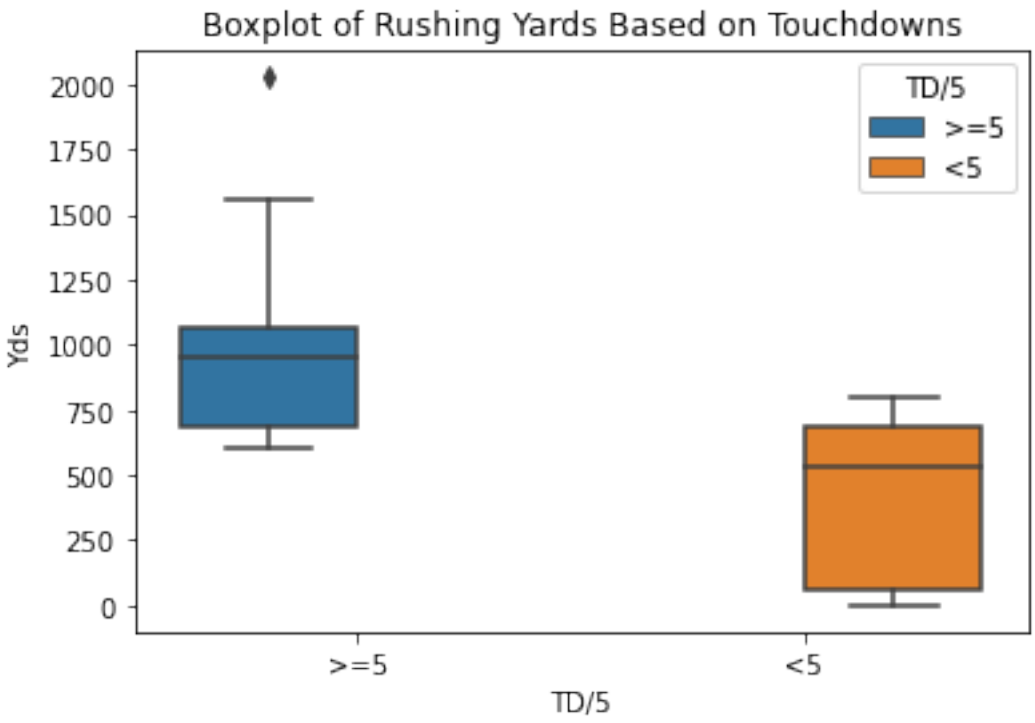
[143]:

		Min	1st Quartile	Median	3rd Quartile	Max
0	Y/G	-0.1	43.35	59.7	71.075	126.7
1	Lng	-1.0	28.75	44.0	62.750	98.0

3.5 Q2(e)

[150]:
```
sb.boxplot(x='TD/5',
           y='Yds',
           hue='TD/5',
           data=rb6[['Yds', 'TD/5']])
plt.title("Boxplot of Rushing Yards Based on Touchdowns")
```

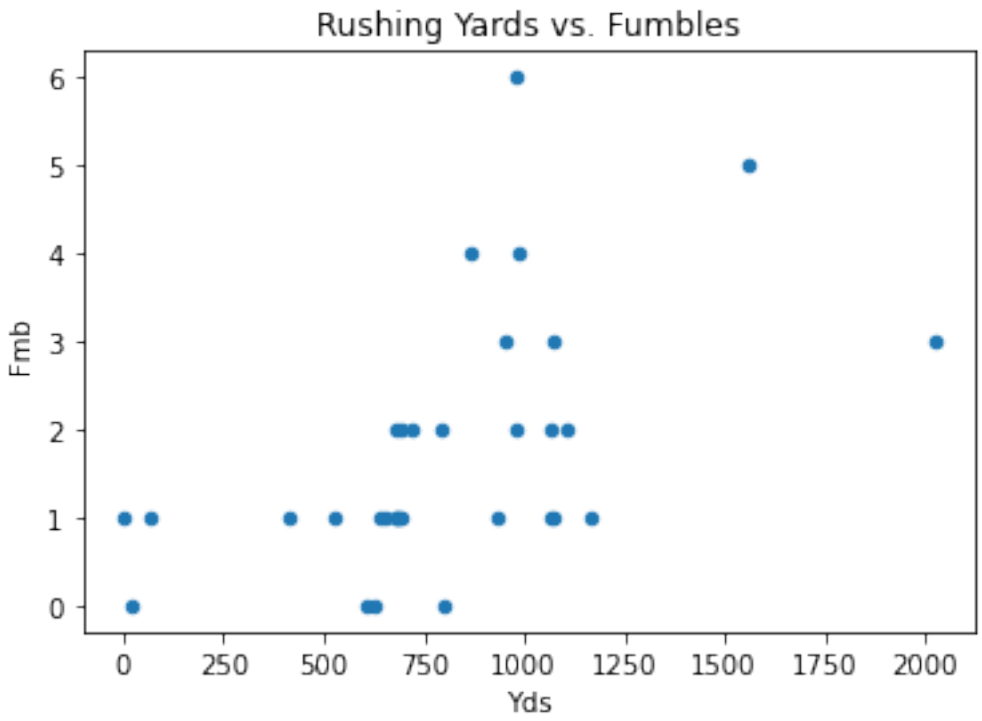
[150]: Text(0.5, 1.0, 'Boxplot of Rushing Yards Based on Touchdowns')



3.6 Q2(f)

[151]:
```
rb6.plot.scatter(x="Yds", y="Fmb")
plt.title("Rushing Yards vs. Fumbles")
```

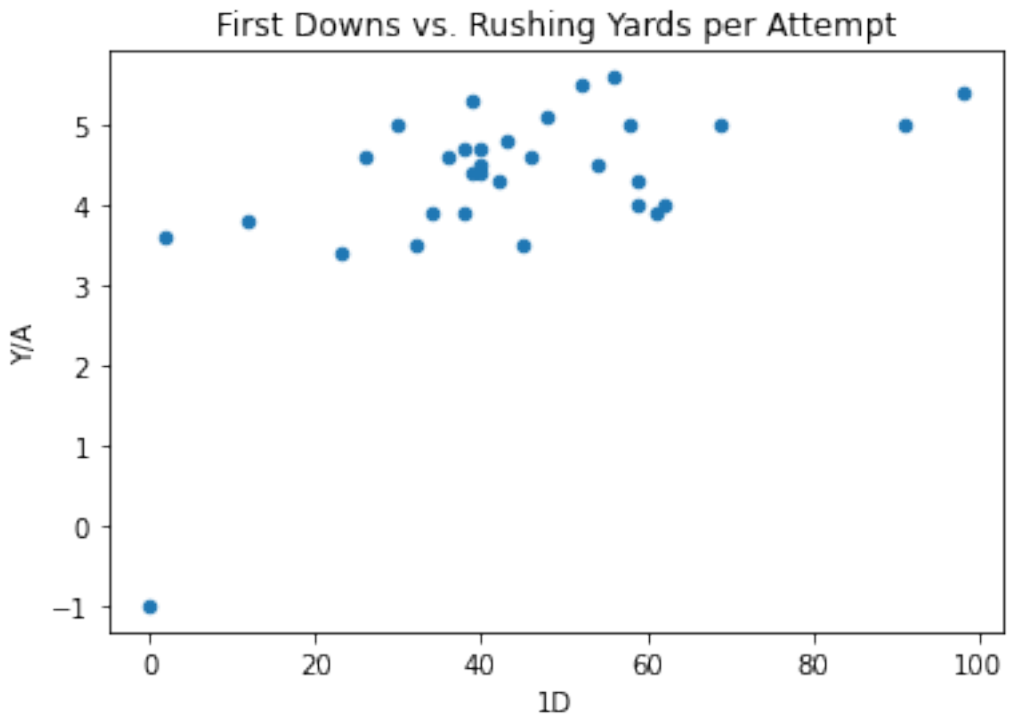
[151]:
```
Text(0.5, 1.0, 'Rushing Yards vs. Fumbles')
```



3.7 Q2(g)

[152]:
```
rb6.plot.scatter(x="1D", y="Y/A")
plt.title("First Downs vs. Rushing Yards per Attempt")
```

[152]:
```
Text(0.5, 1.0, 'First Downs vs. Rushing Yards per Attempt')
```


First Downs vs. Rushing Yards per Attempt
Y/A
1D
5
4
3
2
1
0
−1
0
20
40
60
80
100