In [278…
```python
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

import warnings
warnings.filterwarnings('ignore')
```

In [279…
```python
data= pd.read_csv('train_auto.csv')
```

In [280…
```python
data.head()
```

Out[280…

	INDEX	TARGET_FLAG	TARGET_AMT	KIDSDRIV	AGE	HOMEKIDS	YOJ	INCOME	PARENT1	HOME_VAL	...	BLUEBOOK	TIF	CAR_TYPE
0	1	0	0.0	0	60.0	0	11.0	$67,349	No	$0	...	$14,230	11	Minivan
1	2	0	0.0	0	43.0	0	11.0	$91,449	No	$257,252	...	$14,940	1	Minivan
2	4	0	0.0	0	35.0	1	10.0	$16,039	No	$124,191	...	$4,010	4	z_SUV
3	5	0	0.0	0	51.0	0	14.0	NaN	No	$306,251	...	$15,440	7	Minivan
4	6	0	0.0	0	50.0	0	NaN	$114,986	No	$243,925	...	$18,000	1	z_SUV

5 rows × 26 columns

In [281…
```python
data.tail()
```

Out[281…

	INDEX	TARGET_FLAG	TARGET_AMT	KIDSDRIV	AGE	HOMEKIDS	YOJ	INCOME	PARENT1	HOME_VAL	...	BLUEBOOK	TIF	CAR_TYPE
8156	10297	0	0.0	0	35.0	0	11.0	$43,112	No	$0	...	$27,330	10	Panel Truck
8157	10298	0	0.0	1	45.0	2	9.0	$164,669	No	$386,273	...	$13,270	15	Minivan
8158	10299	0	0.0	0	46.0	0	9.0	$107,204	No	$332,591	...	$24,490	6	Panel Truck
8159	10301	0	0.0	0	50.0	0	7.0	$43,445	No	$149,248	...	$22,550	6	Minivan
8160	10302	0	0.0	0	52.0	0	11.0	$53,235	No	$197,017	...	$19,400	6	Minivan

5 rows × 26 columns

In [282…
```python
data.columns
```

Out[282…
```
Index(['INDEX', 'TARGET_FLAG', 'TARGET_AMT', 'KIDSDRIV', 'AGE', 'HOMEKIDS',
       'YOJ', 'INCOME', 'PARENT1', 'HOME_VAL', 'MSTATUS', 'SEX', 'EDUCATION',
       'JOB', 'TRAVTIME', 'CAR_USE', 'BLUEBOOK', 'TIF', 'CAR_TYPE', 'RED_CAR',
       'OLDCLAIM', 'CLM_FREQ', 'REVOKED', 'MVR_PTS', 'CAR_AGE', 'URBANICITY'],
      dtype='object')
```

In [283…
```python
data.isnull().sum()
```

Out[283…
```
INDEX            0
TARGET_FLAG      0
TARGET_AMT       0
KIDSDRIV         0
AGE              6
HOMEKIDS         0
YOJ            454
INCOME         445
PARENT1          0
HOME_VAL       464
MSTATUS          0
SEX              0
EDUCATION        0
JOB            526
TRAVTIME         0
CAR_USE          0
BLUEBOOK         0
TIF              0
CAR_TYPE         0
RED_CAR          0
OLDCLAIM         0
CLM_FREQ         0
REVOKED          0
MVR_PTS          0
CAR_AGE        510
URBANICITY       0
dtype: int64
```

In [284…
```python
data.shape
```

Out[284…
```
(8161, 26)
```

In [285…
```python
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8161 entries, 0 to 8160
Data columns (total 26 columns):
INDEX          8161 non-null int64
TARGET_FLAG    8161 non-null int64
TARGET_AMT     8161 non-null float64
KIDSDRIV       8161 non-null int64
AGE            8155 non-null float64
HOMEKIDS       8161 non-null int64
YOJ            7707 non-null float64
INCOME         7716 non-null object
PARENT1        8161 non-null object
HOME_VAL       7697 non-null object
MSTATUS        8161 non-null object
SEX            8161 non-null object
EDUCATION      8161 non-null object
JOB            7635 non-null object
TRAVTIME       8161 non-null int64
CAR_USE        8161 non-null object
BLUEBOOK       8161 non-null object
TIF            8161 non-null int64
CAR_TYPE       8161 non-null object
RED_CAR        8161 non-null object
OLDCLAIM       8161 non-null object
CLM_FREQ       8161 non-null int64
REVOKED        8161 non-null object
MVR_PTS        8161 non-null int64
CAR_AGE        7651 non-null float64
URBANICITY     8161 non-null object
dtypes: float64(4), int64(8), object(14)
memory usage: 1.2+ MB
```

In [286…
```python
data.describe()
```

Out[286…

	INDEX	TARGET_FLAG	TARGET_AMT	KIDSDRIV	AGE	HOMEKIDS	YOJ	TRAVTIME	TIF	CLM_FREQ
count	8161.000000	8161.000000	8161.000000	8161.000000	8155.000000	8161.000000	7707.000000	8161.000000	8161.000000	8161.000000
mean	5151.867663	0.263816	1504.324648	0.171057	44.790313	0.721235	10.499286	33.485725	5.351305	0.798554
std	2978.893962	0.440728	4704.026930	0.511534	8.627589	1.116323	4.092474	15.908333	4.146635	1.158453
min	1.000000	0.000000	0.000000	0.000000	16.000000	0.000000	0.000000	5.000000	1.000000	0.000000
25%	2559.000000	0.000000	0.000000	0.000000	39.000000	0.000000	9.000000	22.000000	1.000000	0.000000
50%	5133.000000	0.000000	0.000000	0.000000	45.000000	0.000000	11.000000	33.000000	4.000000	0.000000
75%	7745.000000	1.000000	1036.000000	0.000000	51.000000	1.000000	13.000000	44.000000	7.000000	2.000000
max	10302.000000	1.000000	107586.136160	4.000000	81.000000	5.000000	23.000000	142.000000	25.000000	5.000000

In [287…
```python
data.duplicated().sum()
```

Out[287…
```
0
```

In [288…
```python
data.nunique()
```

Out[288…
```
INDEX          8161
TARGET_FLAG       2
TARGET_AMT     1949
KIDSDRIV          5
AGE              60
HOMEKIDS          6
YOJ              21
INCOME         6612
PARENT1           2
HOME_VAL       5106
MSTATUS           2
SEX               2
EDUCATION         5
JOB               8
TRAVTIME         97
CAR_USE           2
BLUEBOOK       2789
TIF              23
CAR_TYPE          6
RED_CAR           2
OLDCLAIM       2857
CLM_FREQ          6
REVOKED           2
MVR_PTS          13
CAR_AGE          30
URBANICITY        2
dtype: int64
```

In [289…
```python
data['INCOME'].head()
```

Out[289…
```
0     $67,349
1     $91,449
2     $16,039
3         NaN
4    $114,986
Name: INCOME, dtype: object
```

In [290…
```python
data['INCOME'] = data['INCOME'].apply(lambda x :str(x).replace('$',''))
```

In [291…
```python
data['INCOME'] = data['INCOME'].apply(lambda x :str(x).replace(',',''))
```

In [292…
```python
data['INCOME']=  data['INCOME'].apply(lambda x :str(x).replace(' ',''))
```

In [293…
```python
data['INCOME']=pd.to_numeric(data['INCOME'],errors='coerce')
```

In [294…
```python
data['INCOME'].mean()
```

Out[294…
```
61898.0946086055
```

In [295…
```python
data['INCOME'].median()
```

Out[295…
```
54028.0
```

In [296…
```python
data.INCOME.isnull().sum()/len(data.index)*100
```

Out[296…
```
5.452763141771841
```

In [297…
```python
data['INCOME'].fillna(data['INCOME'].median(),inplace=True)
```

In [298…
```python
data['HOME_VAL'].head()
```

Out[298…
```
0          $0
1    $257,252
2    $124,191
3    $306,251
4    $243,925
Name: HOME_VAL, dtype: object
```

In [299…
```python
data['HOME_VAL'] = data['HOME_VAL'].apply(lambda x:str(x).replace('$',''))
```

In [300…
```python
data['HOME_VAL'] = data['HOME_VAL'].apply(lambda x:str(x).replace(',',''))
```

In [301…
```python
data['HOME_VAL'].isnull().sum()
```

Out[301…
```
0
```

In [302…
```python
data['HOME_VAL']=pd.to_numeric(data['HOME_VAL'],errors='coerce')
```

In [303…
```python
data['HOME_VAL'].mean()
```

Out[303…
```
154867.2897232688
```

In [304…
```python
data['HOME_VAL'].median()
```

Out[304…
```
161160.0
```

In [305…
```python
data['BLUEBOOK']=data['BLUEBOOK'].apply(lambda x:str(x).replace('$',''))
data['BLUEBOOK']=data['BLUEBOOK'].apply(lambda x:str(x).replace(',',''))
data['BLUEBOOK']=pd.to_numeric(data['BLUEBOOK'],errors='coerce')
```

In [306…
```python
data['BLUEBOOK'].mean()
```

Out[306…
```
15709.899522117388
```

In [307…
```python
data['BLUEBOOK'].median()
```

Out[307…
```
14440.0
```

In [308…
```python
data.dtypes
```

Out[308…
```
INDEX            int64
TARGET_FLAG      int64
TARGET_AMT     float64
KIDSDRIV         int64
AGE            float64
HOMEKIDS         int64
YOJ            float64
INCOME         float64
PARENT1         object
HOME_VAL       float64
MSTATUS         object
SEX             object
EDUCATION       object
JOB             object
TRAVTIME         int64
CAR_USE         object
BLUEBOOK         int64
TIF              int64
CAR_TYPE        object
RED_CAR         object
OLDCLAIM        object
CLM_FREQ         int64
REVOKED         object
MVR_PTS          int64
CAR_AGE        float64
URBANICITY      object
dtype: object
```

In [309…
```python
data['OLDCLAIM'].head()
```

Out[309…
```
0     $4,461
1         $0
2    $38,690
3         $0
4    $19,217
Name: OLDCLAIM, dtype: object
```

In [310…
```python
data['OLDCLAIM']= data['OLDCLAIM'].apply(lambda x:str(x).replace('$',''))
data['OLDCLAIM']=data['OLDCLAIM'].apply(lambda x:str(x).replace(',',''))
data['OLDCLAIM']=pd.to_numeric(data['OLDCLAIM'],errors='coerce')
```

In [311…
```python
data['OLDCLAIM'].mean()
```

Out[311…
```
4037.0762161499815
```

In [312…
```python
data['OLDCLAIM'].median()
```

Out[312…
```
0.0
```

In [313…
```python
data['YOJ'].isnull().sum()/len(data.index)*100
```

Out[313…
```
5.563043744639137
```

In [314…
```python
data['YOJ'].mean()
```

Out[314…
```
10.499286363046581
```

In [315…
```python
b=data['YOJ'].median()
```

In [316…
```python
data['YOJ'].fillna(b,inplace=True)
```

In [317…
```python
data.isnull().sum()
```

Out[317…
```
INDEX            0
TARGET_FLAG      0
TARGET_AMT       0
KIDSDRIV         0
AGE              6
HOMEKIDS         0
YOJ              0
INCOME           0
PARENT1          0
HOME_VAL       464
MSTATUS          0
SEX              0
EDUCATION        0
JOB            526
TRAVTIME         0
CAR_USE          0
BLUEBOOK         0
TIF              0
CAR_TYPE         0
RED_CAR          0
OLDCLAIM         0
CLM_FREQ         0
REVOKED          0
MVR_PTS          0
CAR_AGE        510
URBANICITY       0
dtype: int64
```

In [318…
```python
data.HOME_VAL.isnull().sum()/len(data.index)*100
```

Out[318…
```
5.685577747825022
```

In [319…
```python
data['HOME_VAL'].mean()
```

Out[319…
```
154867.2897232688
```

In [320…
```python
data['HOME_VAL'].median()
```

Out[320…
```
161160.0
```

In [321…
```python
data['HOME_VAL'].fillna(data['HOME_VAL'].mean(),inplace=True)
```

In [322…
```python
data['JOB'].value_counts()
```

Out[322…
```
z_Blue Collar    1825
Clerical         1271
Professional     1117
Manager           988
Lawyer            835
Student           712
Home Maker        641
Doctor            246
Name: JOB, dtype: int64
```

In [323…
```python
c = 'z_Blue Collar'
data['JOB'].fillna(c,inplace=True)
```

In [324…
```python
data['CAR_AGE'].isnull().sum()/len(data.index)*100
```

Out[324…
```
6.2492341624800885
```

In [325…
```python
data['CAR_AGE'].mean()
```

Out[325…
```
8.32832309502026
```

In [326…
```python
v=data['CAR_AGE'].median()
```

In [327…
```python
data['CAR_AGE'].fillna(v,inplace=True)
```

In [328…
```python
data['AGE'].head()
```

Out[328…
```
0    60.0
1    43.0
2    35.0
3    51.0
4    50.0
Name: AGE, dtype: float64
```

In [329…
```python
data['AGE'].mean()
```

Out[329…
```
44.79031269160024
```

In [330…
```python
data['AGE'].median()
```

Out[330…
```
45.0
```

In [331…
```python
data['AGE'].fillna(data['AGE'].mean(),inplace=True)
```

In [332…
```python
data.isnull().sum()
```

Out[332…
```
INDEX          0
TARGET_FLAG    0
TARGET_AMT     0
KIDSDRIV       0
AGE            0
HOMEKIDS       0
YOJ            0
INCOME         0
PARENT1        0
HOME_VAL       0
MSTATUS        0
SEX            0
EDUCATION      0
JOB            0
TRAVTIME       0
CAR_USE        0
BLUEBOOK       0
TIF            0
CAR_TYPE       0
RED_CAR        0
OLDCLAIM       0
CLM_FREQ       0
REVOKED        0
MVR_PTS        0
CAR_AGE        0
URBANICITY     0
dtype: int64
```

In [333…
```python
data.iloc[:,:10].head(10)
```

Out[333…

	INDEX	TARGET_FLAG	TARGET_AMT	KIDSDRIV	AGE	HOMEKIDS	YOJ	INCOME	PARENT1	HOME_VAL
0	1	0	0.0	0	60.0	0	11.0	67349.0	No	0.000000
1	2	0	0.0	0	43.0	0	11.0	91449.0	No	257252.000000
2	4	0	0.0	0	35.0	1	10.0	16039.0	No	124191.000000
3	5	0	0.0	0	51.0	0	14.0	54028.0	No	306251.000000
4	6	0	0.0	0	50.0	0	11.0	114986.0	No	243925.000000
5	7	1	2946.0	0	34.0	1	12.0	125301.0	Yes	0.000000
6	8	0	0.0	0	54.0	0	11.0	18755.0	No	154867.289723
7	11	1	4021.0	1	37.0	2	11.0	107961.0	No	333680.000000
8	12	1	2501.0	0	34.0	0	10.0	62978.0	No	0.000000
9	13	0	0.0	0	50.0	0	7.0	106952.0	No	0.000000

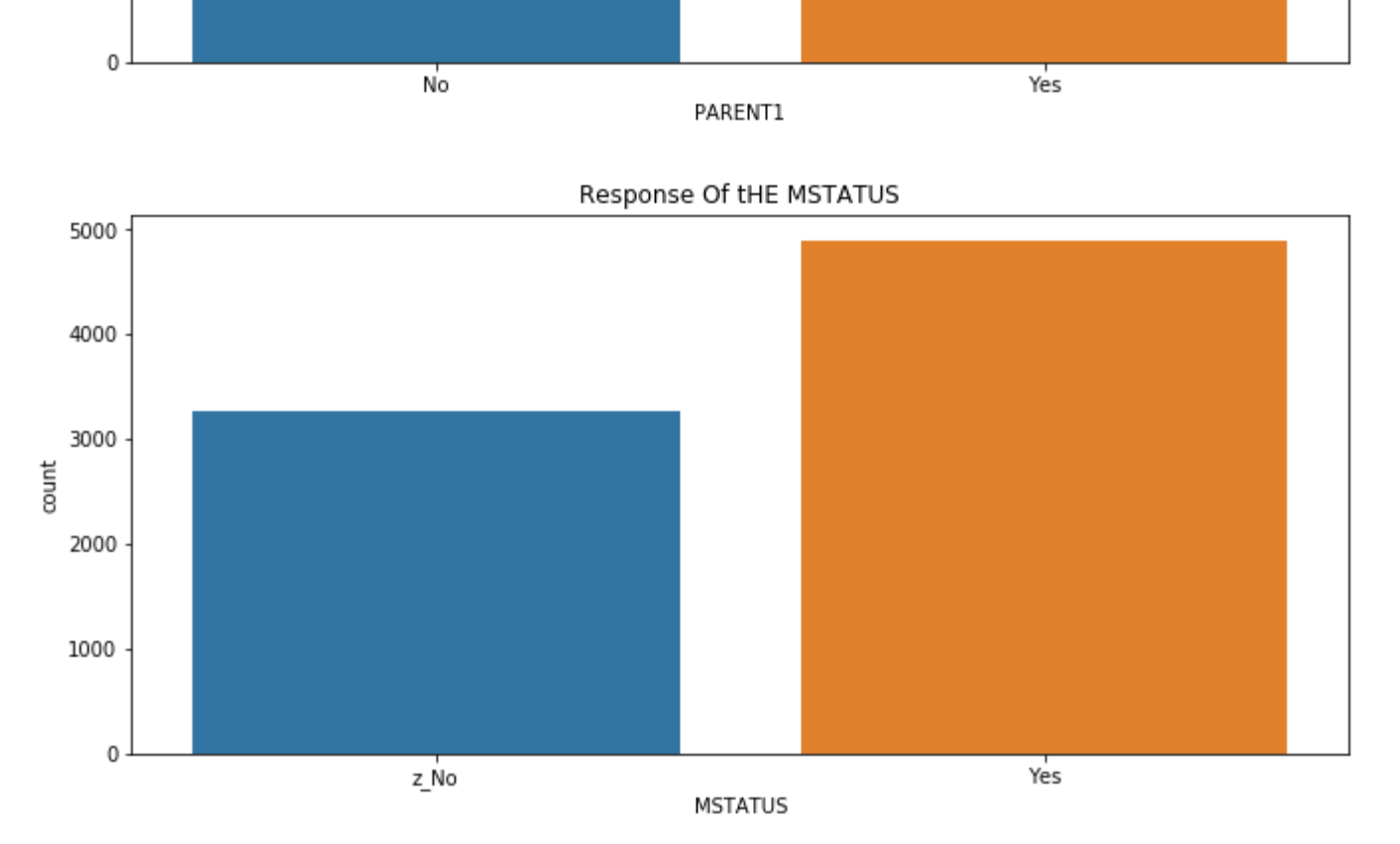
EDA

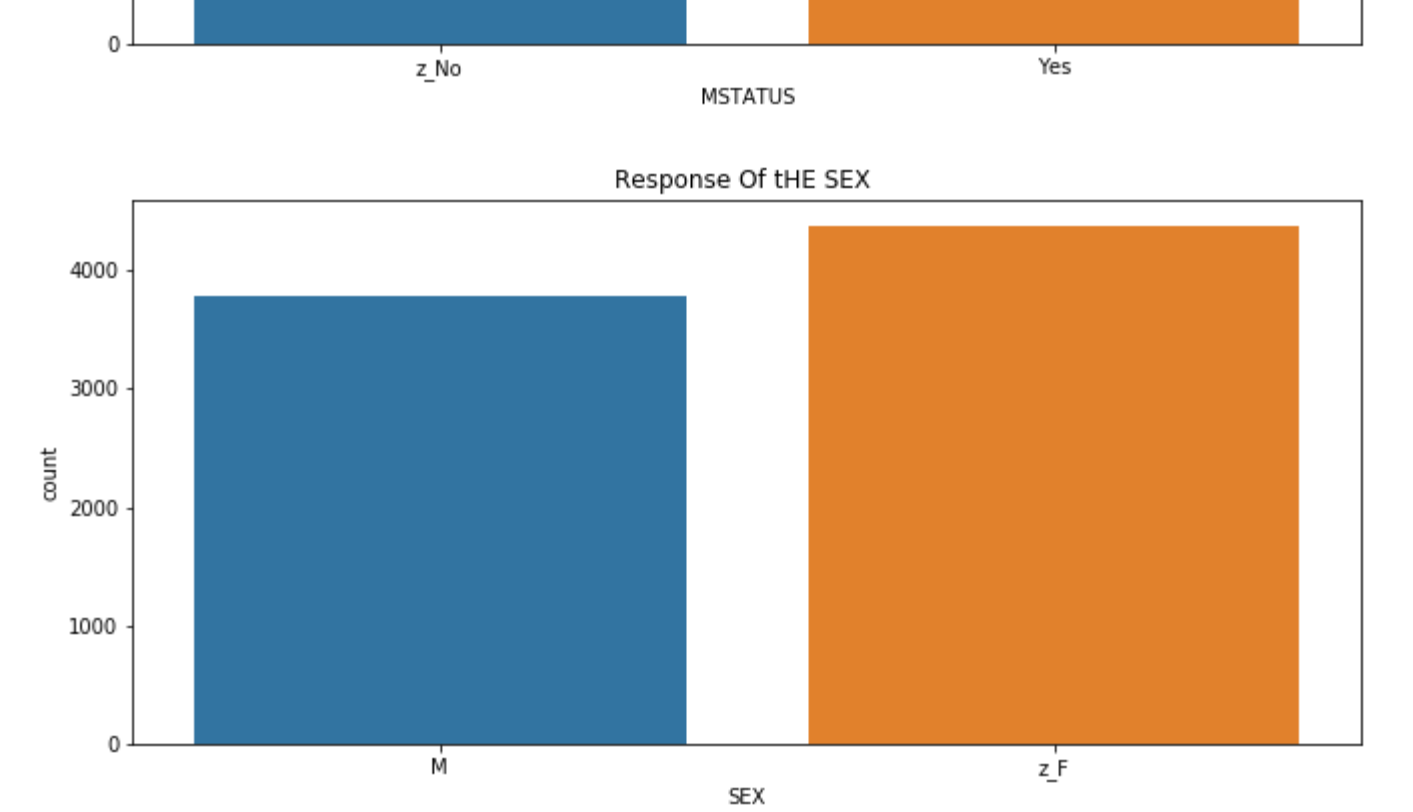
In [334…
```python
plt.figure(figsize=(20,8))
plt.bar(data['AGE'],data['INCOME'])
plt.hist(data['AGE'],density=True,color='teal')
plt.title('AGE VS INcome')
plt.show()
```

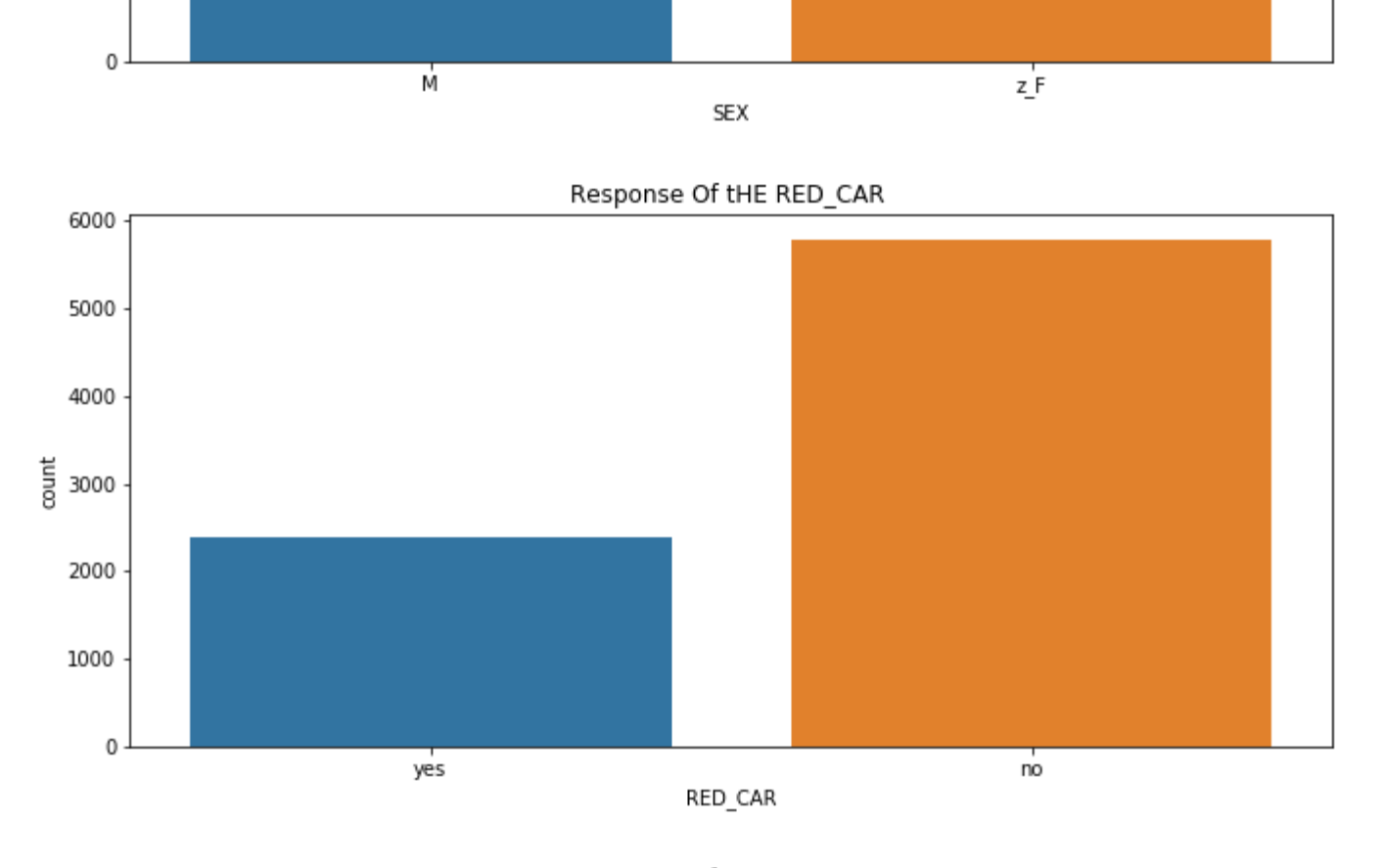
In [335…
```python
data['CAR_TYPE'].value_counts()
```

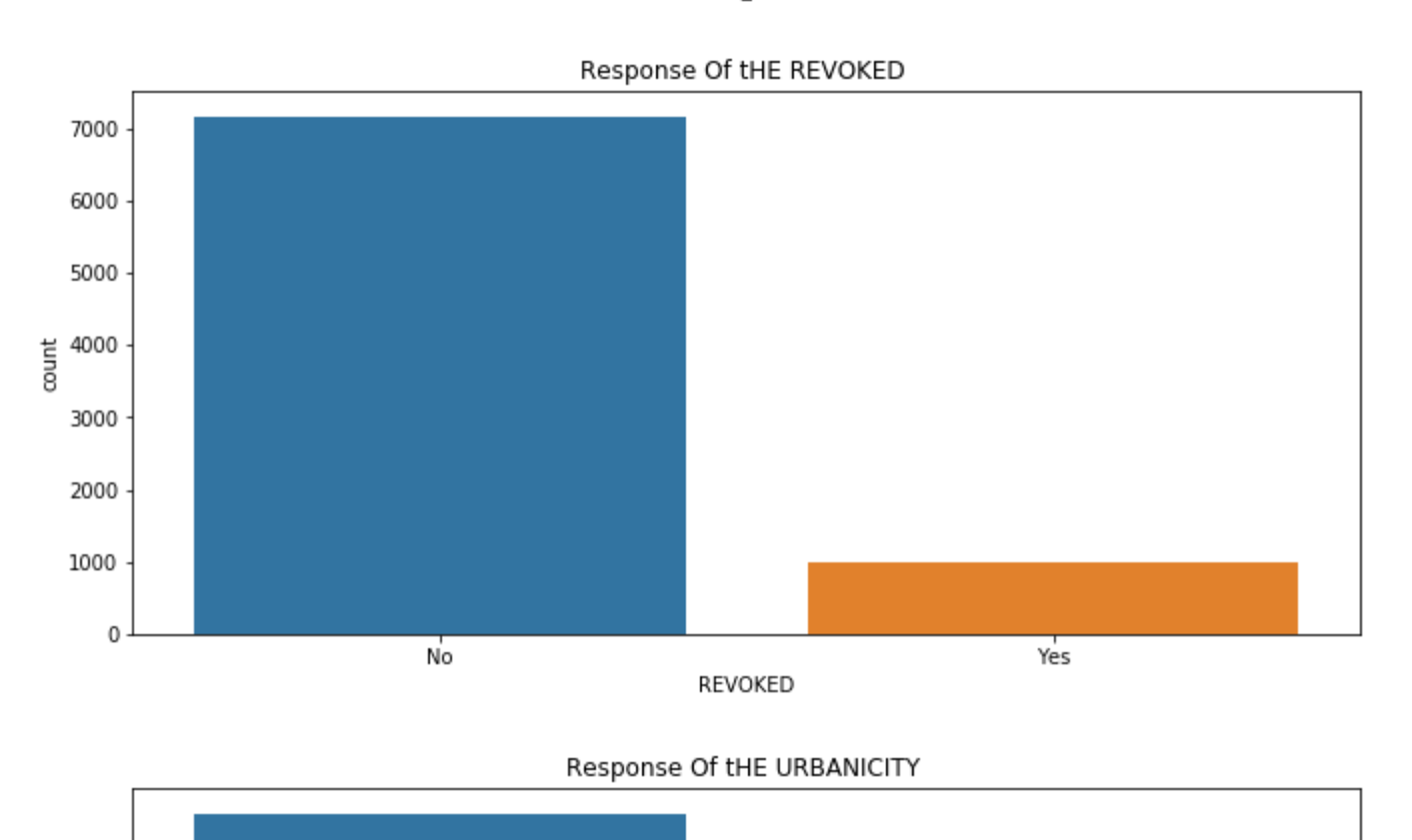
Out[335…
```
z_SUV          2294
Minivan        2145
Pickup         1389
Sports Car      907
Van             750
Panel Truck     676
Name: CAR_TYPE, dtype: int64
```

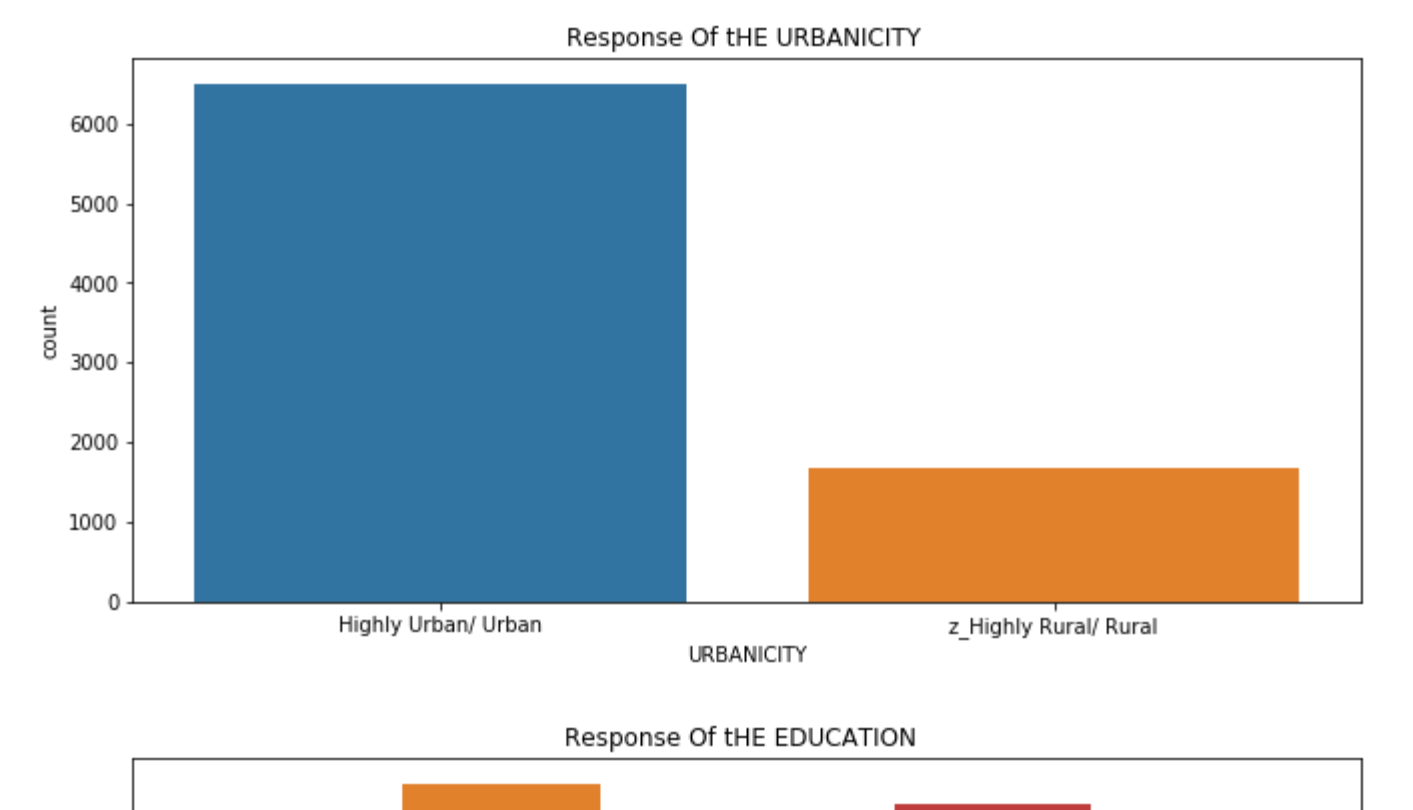
In [336…
```python
fig,ax = plt.subplots(figsize=(11,5))
sns.countplot(data['CAR_TYPE'],ax=ax)
plt.title('Response Of tHE Car Type')
plt.show()
print()
fig,ax = plt.subplots(figsize=(11,5))
sns.countplot(data['PARENT1'],ax=ax)
plt.title('Response Of tHE PARENT1')
plt.show()
print()
fig,ax = plt.subplots(figsize=(11,5))
sns.countplot(data['MSTATUS'],ax=ax)
plt.title('Response Of tHE MSTATUS')
plt.show()
print()
fig,ax = plt.subplots(figsize=(11,5))
sns.countplot(data['SEX'],ax=ax)
plt.title('Response Of tHE SEX ')
plt.show()
print()
fig,ax = plt.subplots(figsize=(11,5))
sns.countplot(data['RED_CAR'],ax=ax)
plt.title('Response Of tHE RED_CAR ')
plt.show()
print()
fig,ax = plt.subplots(figsize=(11,5))
sns.countplot(data['REVOKED'],ax=ax)
plt.title('Response Of tHE REVOKED ')
plt.show()
print()
fig,ax = plt.subplots(figsize=(11,5))
sns.countplot(data['URBANICITY'],ax=ax)
plt.title('Response Of tHE URBANICITY ')
plt.show()
print()
fig,ax = plt.subplots(figsize=(11,5))
sns.countplot(data['EDUCATION'],ax=ax)
plt.title('Response Of tHE EDUCATION  ')
plt.show()
print()
fig,ax = plt.subplots(figsize=(11,5))
sns.countplot(data['JOB'],ax=ax)
plt.title('Response Of tHE JOB')
plt.show()
print()
fig,ax = plt.subplots(figsize=(11,5))
sns.countplot(data['CAR_USE'],ax=ax)
plt.title('Response Of tHE CAR_USE')
plt.show()
```

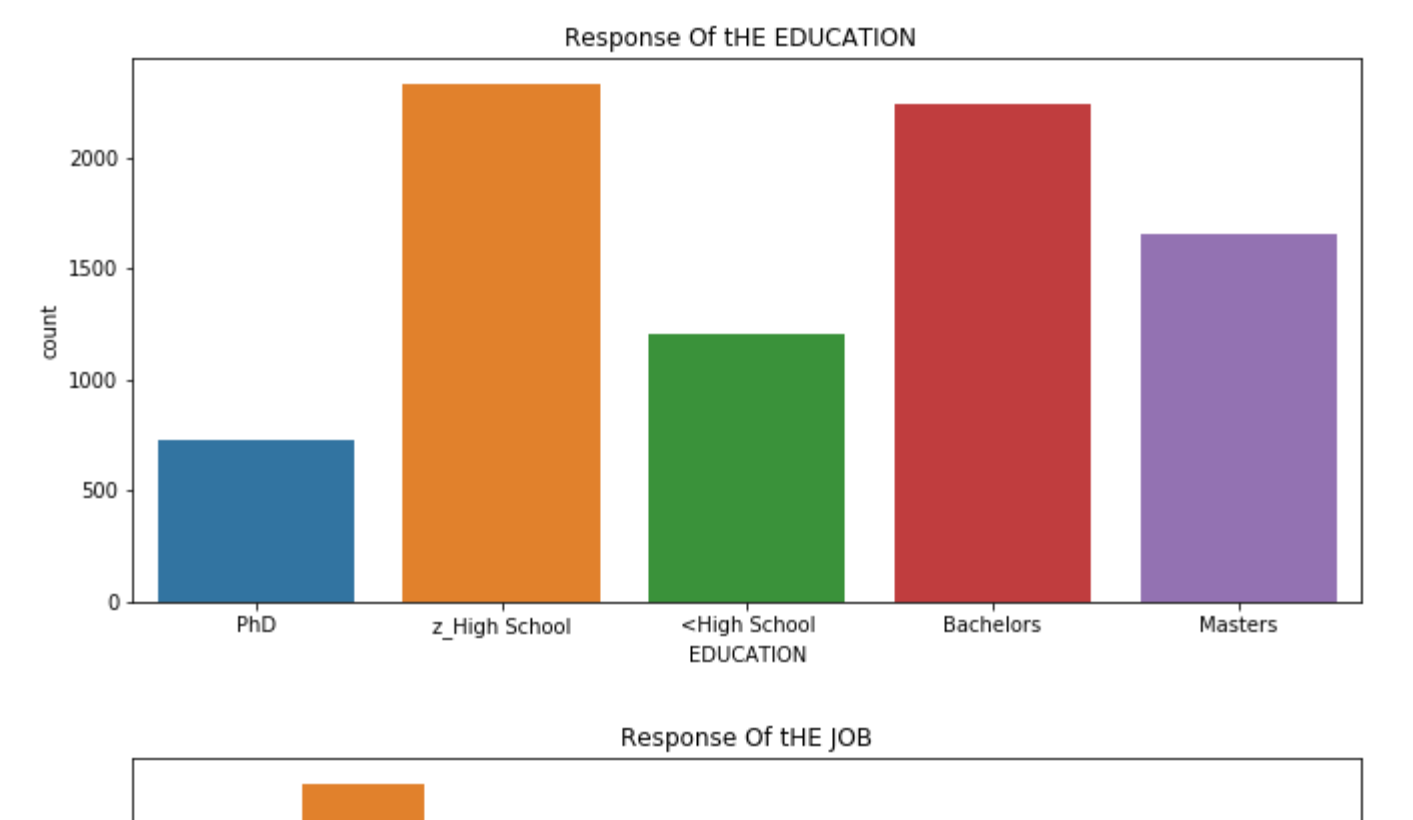


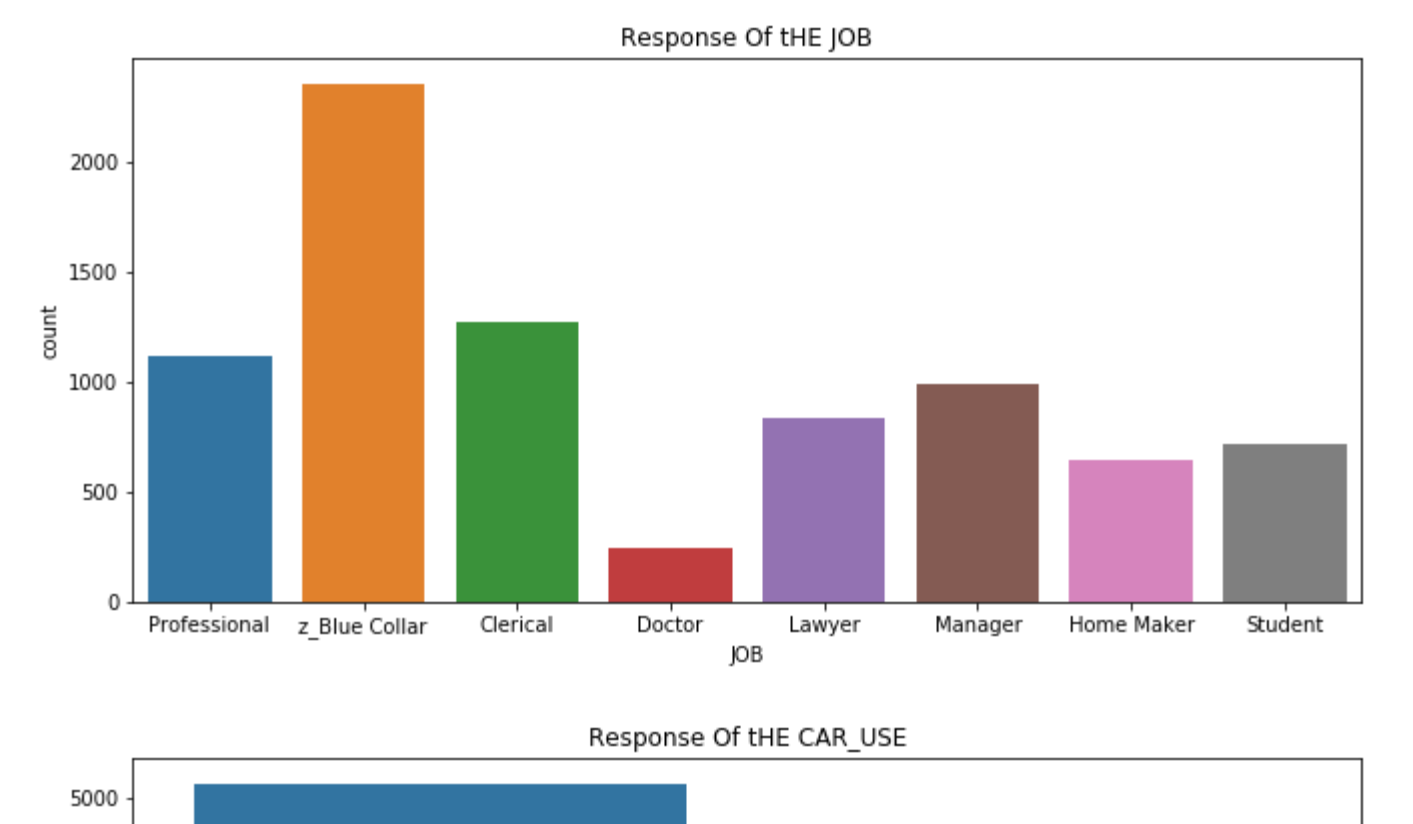


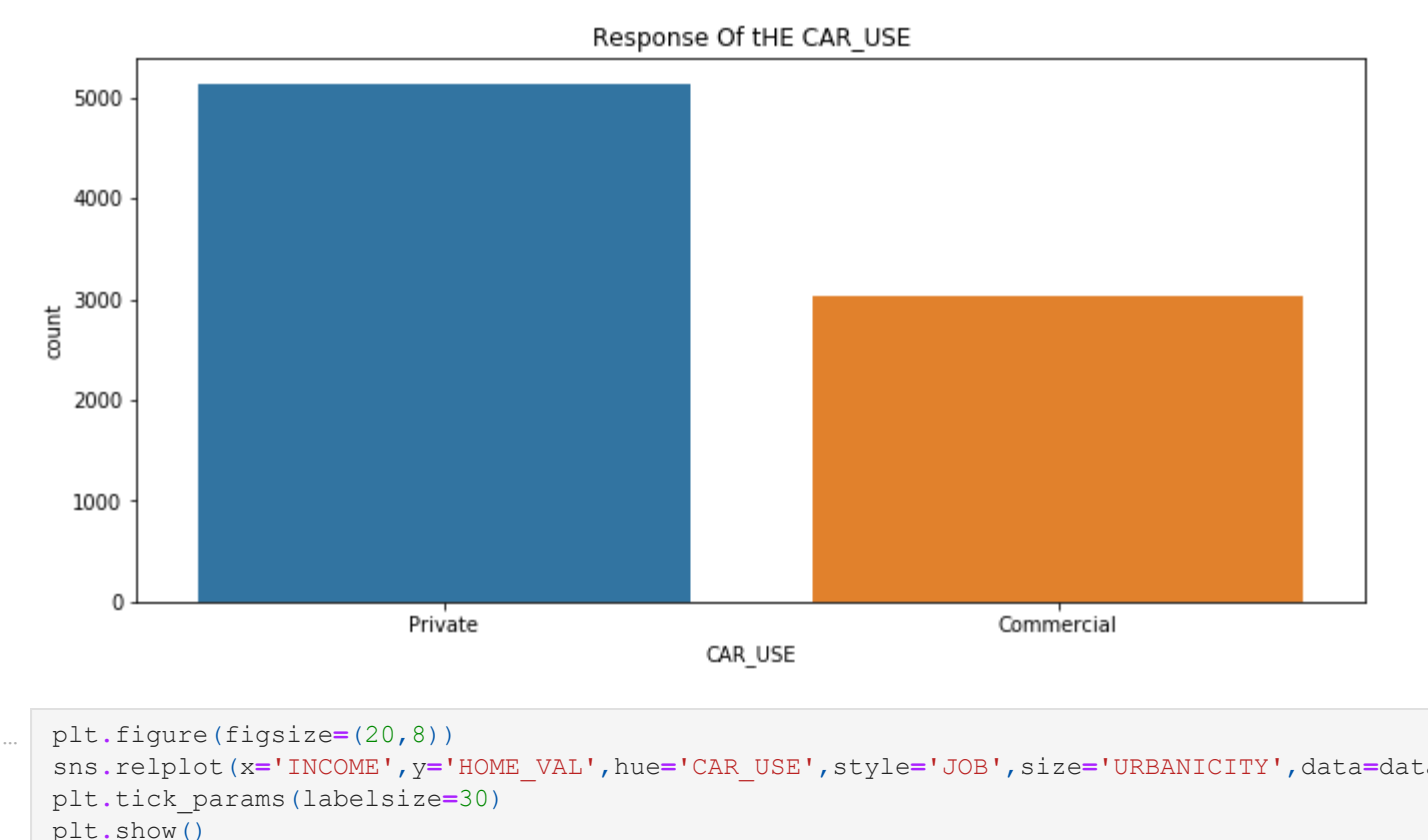






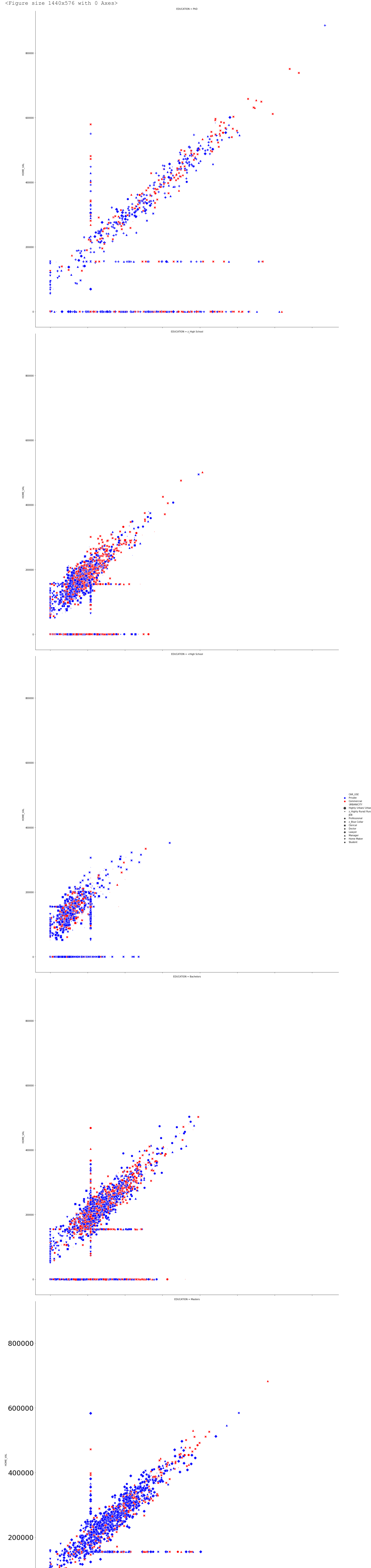




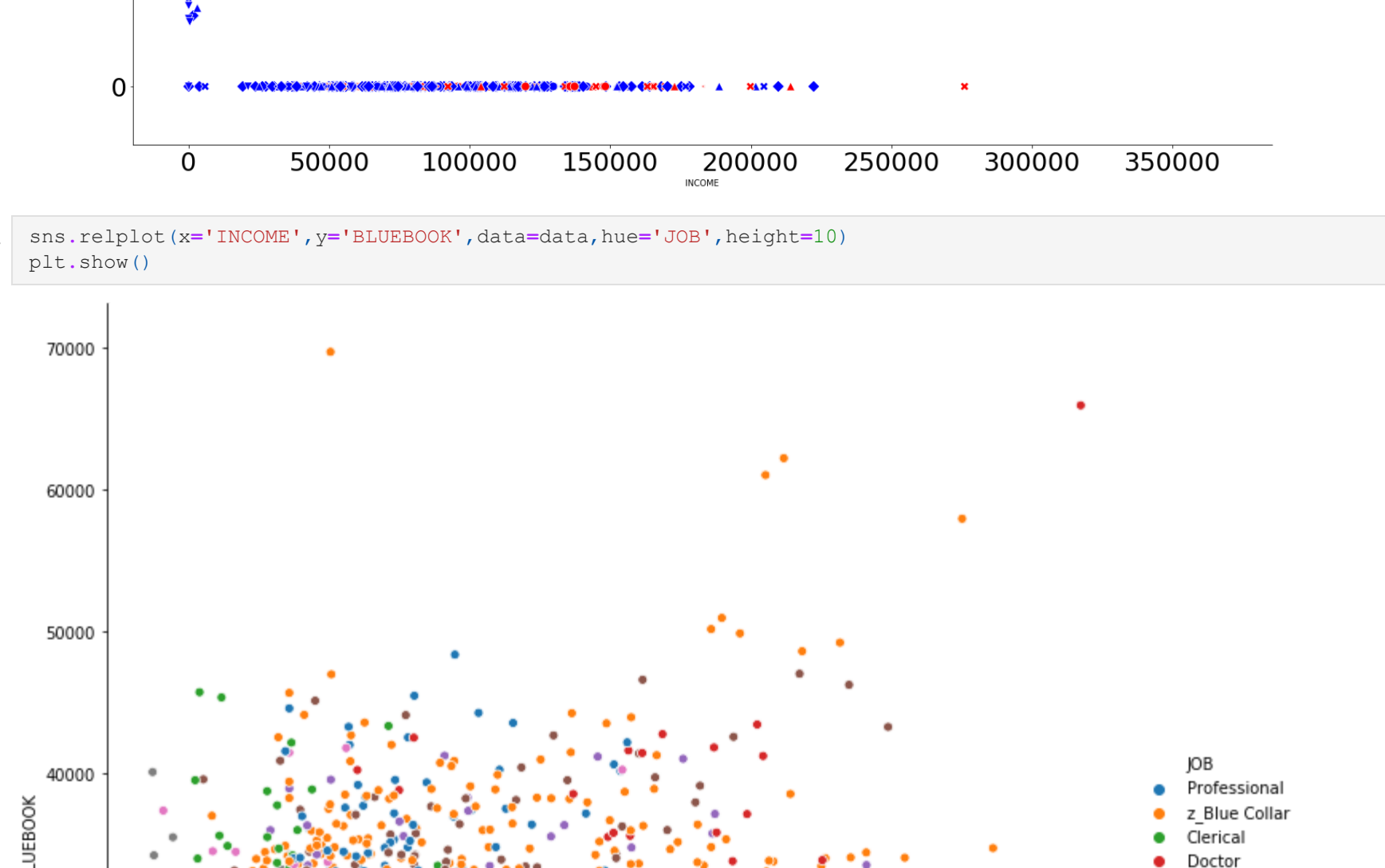


In [337…
```python
plt.figure(figsize=(20,8))
sns.relplot(x='INCOME',y='HOME_VAL',hue='CAR_USE',style='JOB',size='URBANICITY',data=data,palette=['b','r'],si
plt.tick_params(labelsize=30)
plt.show()
```

<Figure size 1440x576 with 0 Axes>

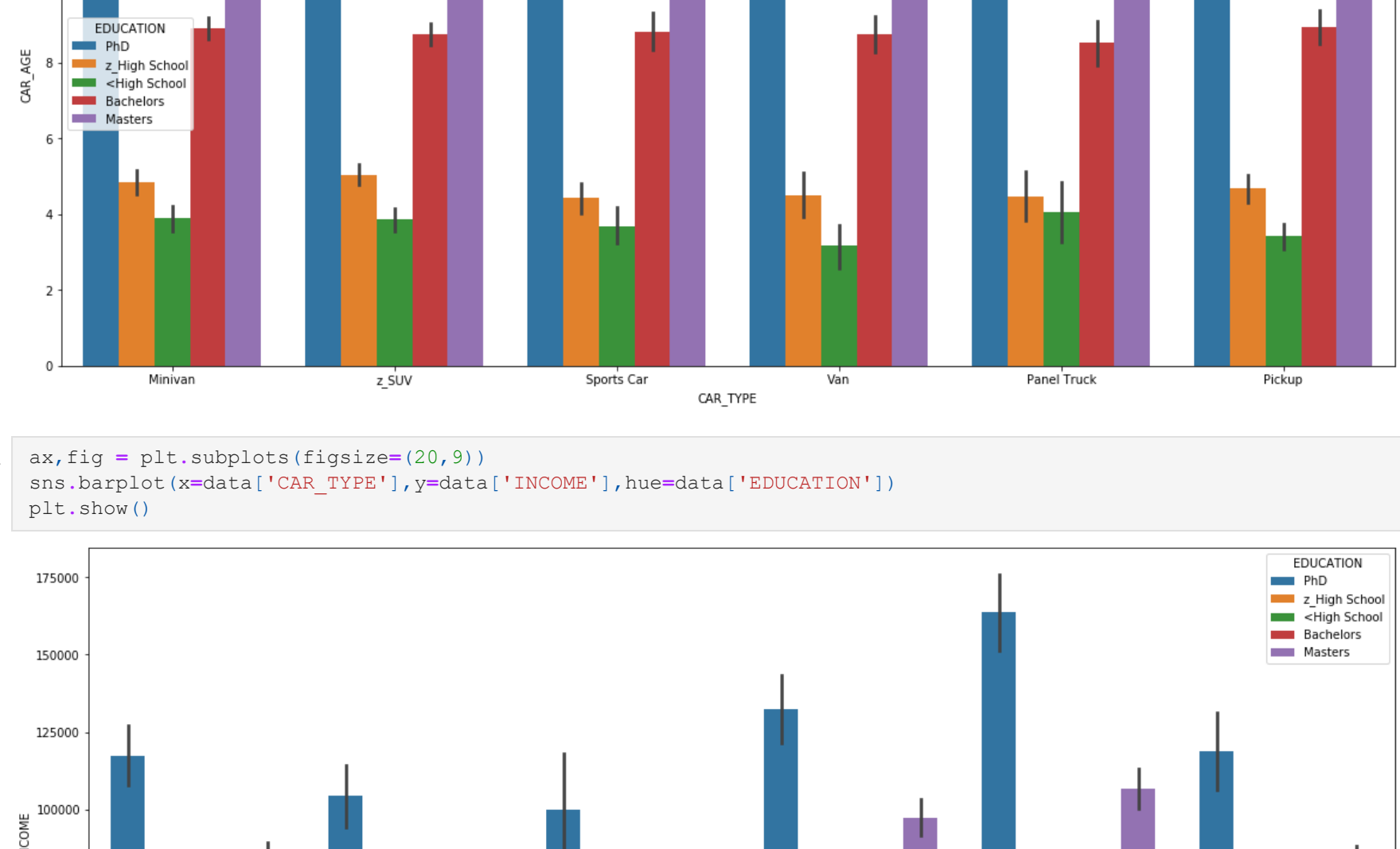


In [338…
```
sns.relplot(x='INCOME',y='BLUEBOOK',data=data,hue='JOB',height=10)
plt.show()
```



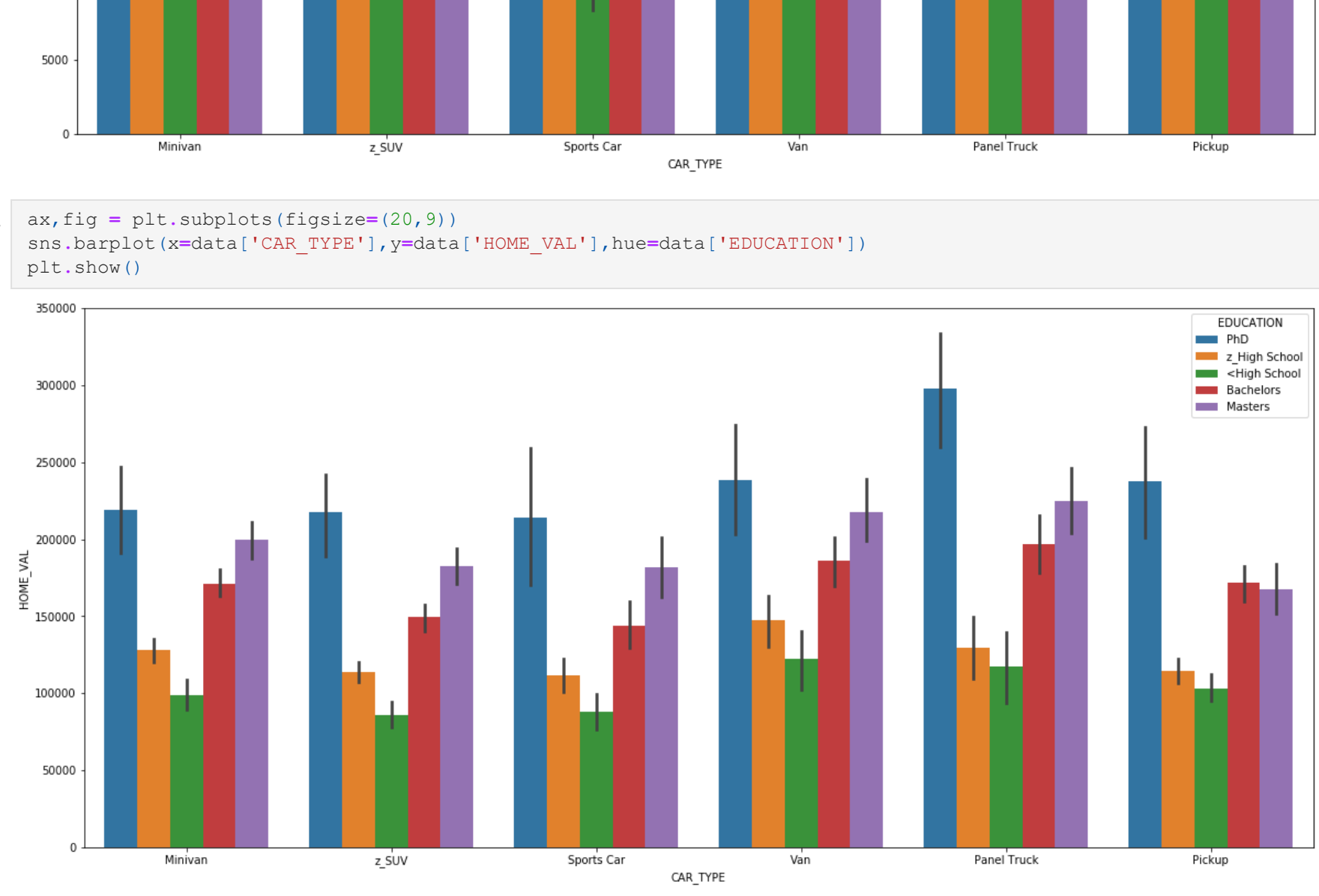
In [339…
```
ax,fig = plt.subplots(figsize=(20,9))
sns.barplot(x=data['CAR_TYPE'],y=data['CAR_AGE'],hue=data['EDUCATION'])
plt.show()
```

In [340…
```
ax,fig = plt.subplots(figsize=(20,9))
sns.barplot(x=data['CAR_TYPE'],y=data['INCOME'],hue=data['EDUCATION'])
plt.show()
```

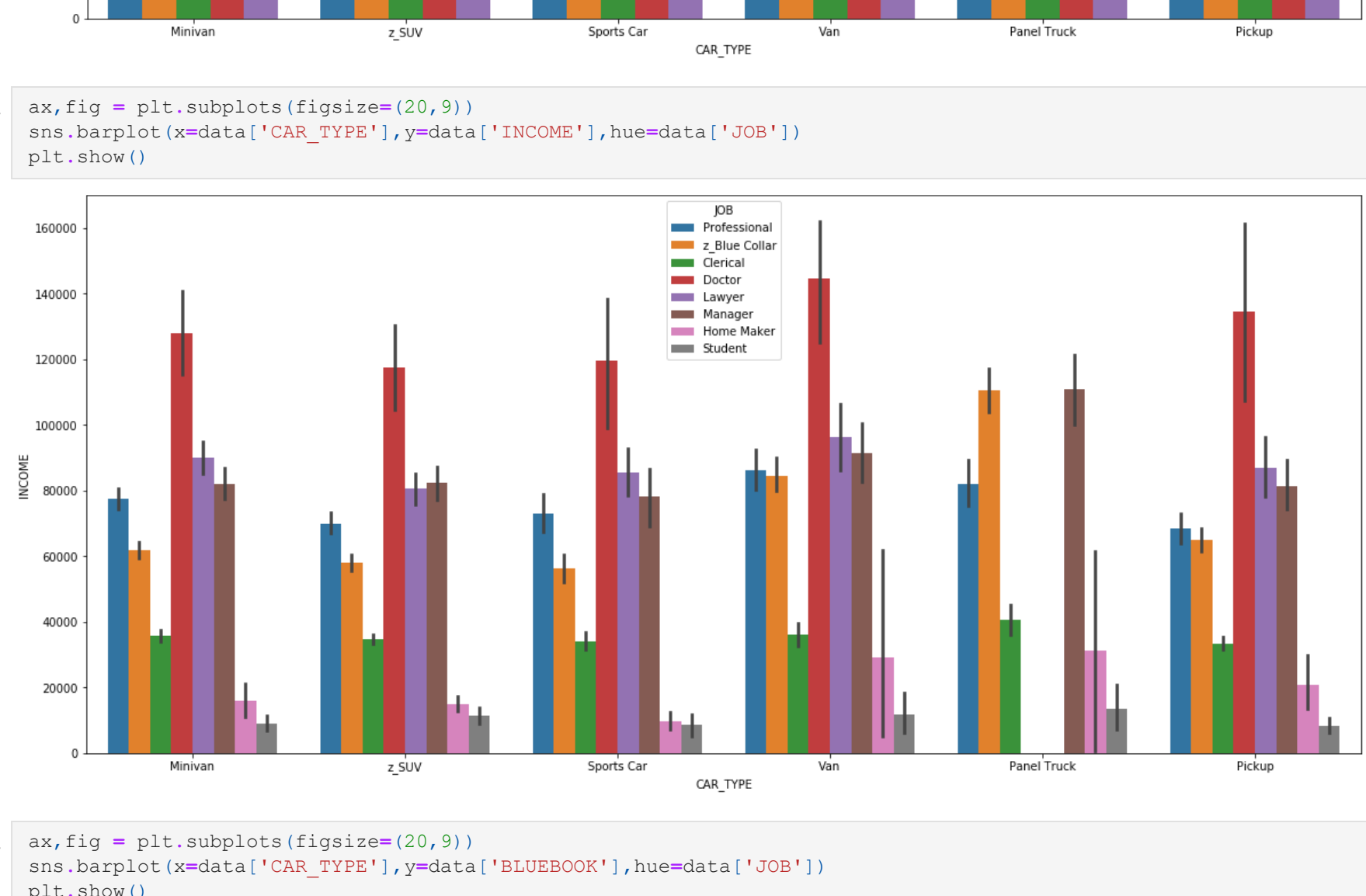


In [341…
```
ax,fig = plt.subplots(figsize=(20,9))
sns.barplot(x=data['CAR_TYPE'],y=data['BLUEBOOK'],hue=data['EDUCATION'])
plt.show()
```

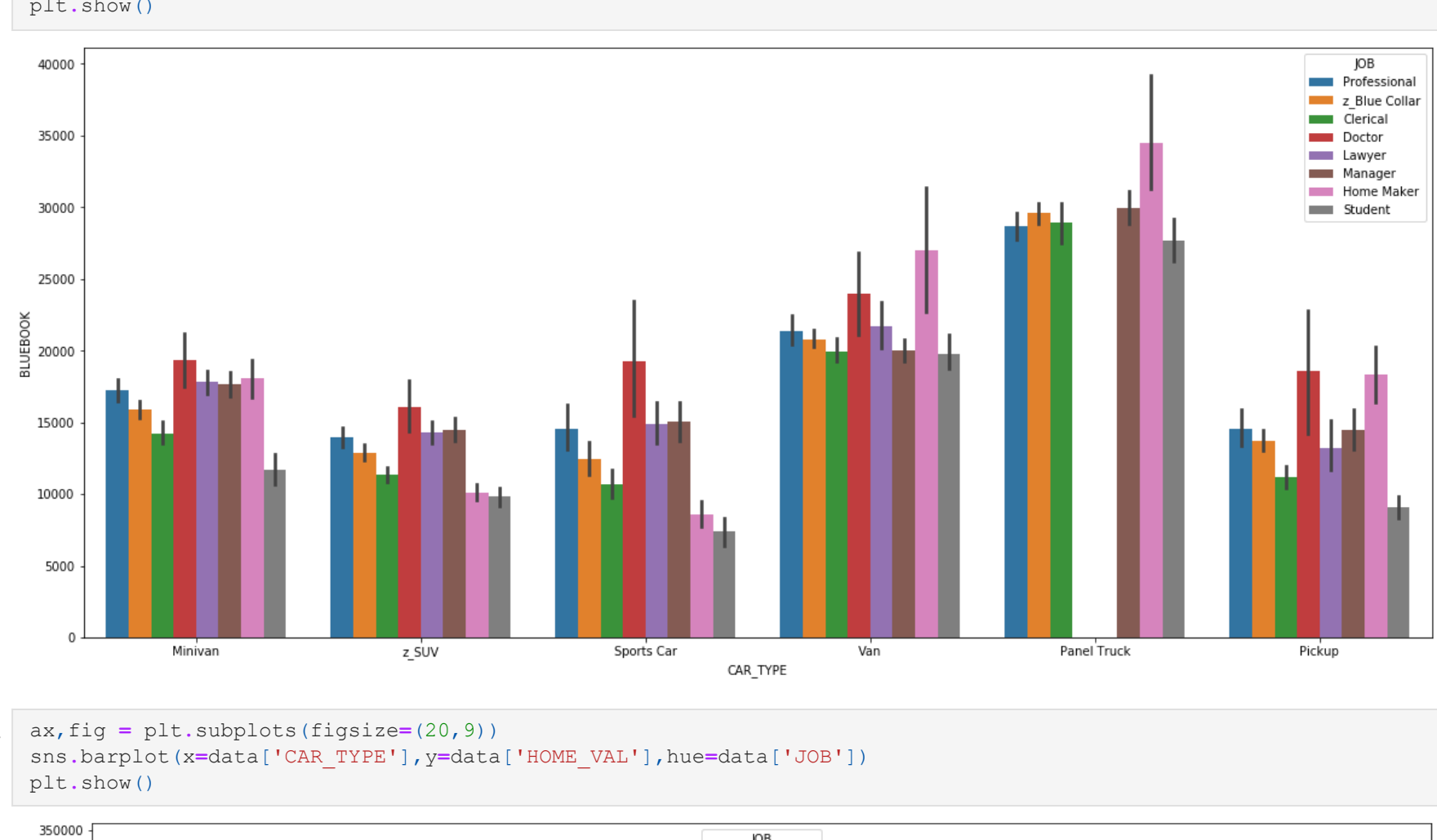
In [342…
```
ax,fig = plt.subplots(figsize=(20,9))
sns.barplot(x=data['CAR_TYPE'],y=data['HOME_VAL'],hue=data['EDUCATION'])
plt.show()
```



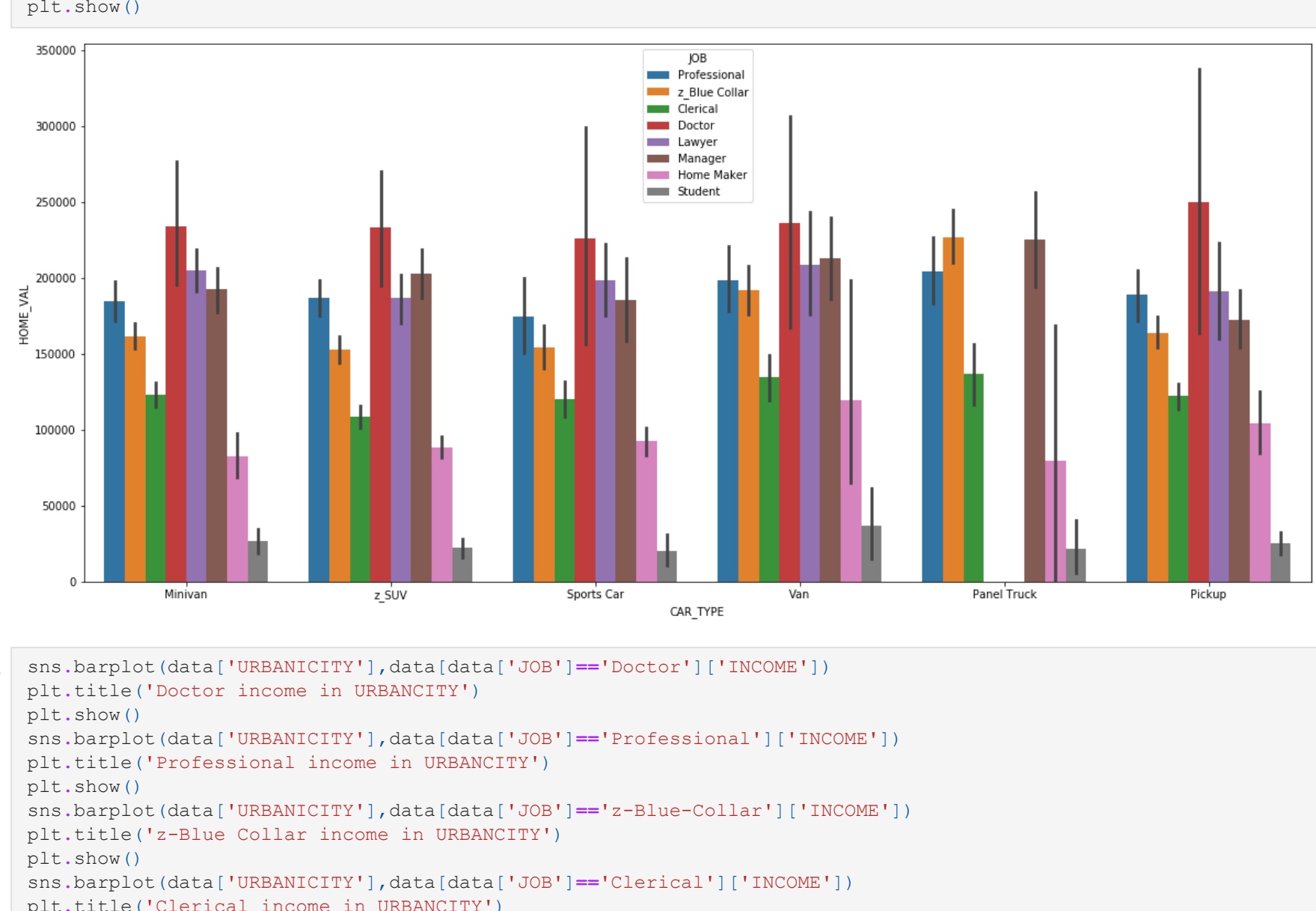
In [343…
```
ax,fig = plt.subplots(figsize=(20,9))
sns.barplot(x=data['CAR_TYPE'],y=data['INCOME'],hue=data['JOB'])
plt.show()
```



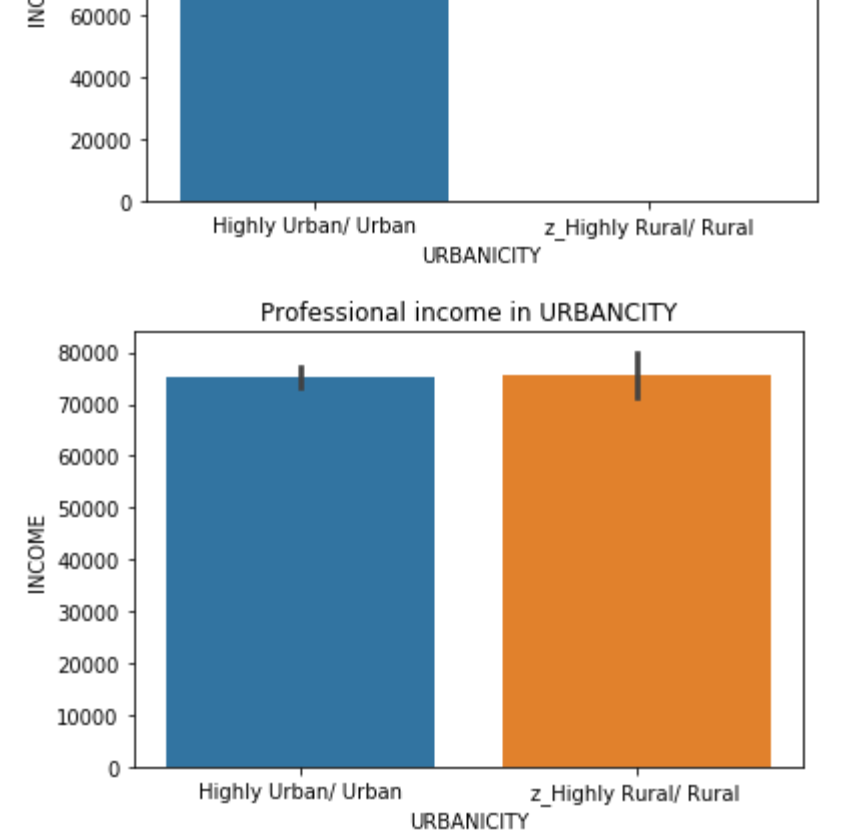
In [344…
```
ax,fig = plt.subplots(figsize=(20,9))
sns.barplot(x=data['CAR_TYPE'],y=data['BLUEBOOK'],hue=data['JOB'])
plt.show()
```

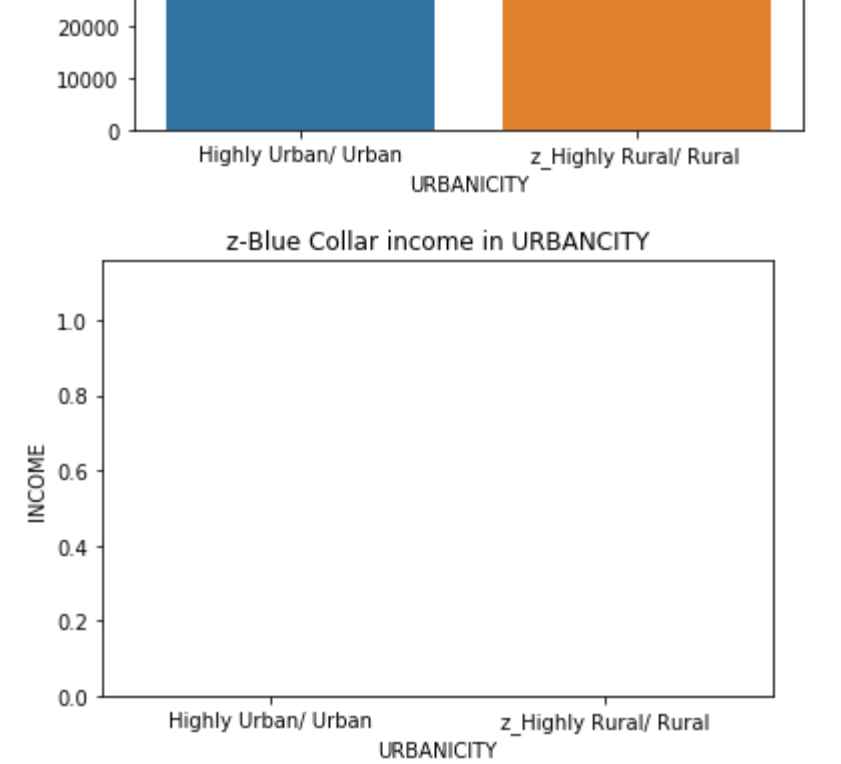


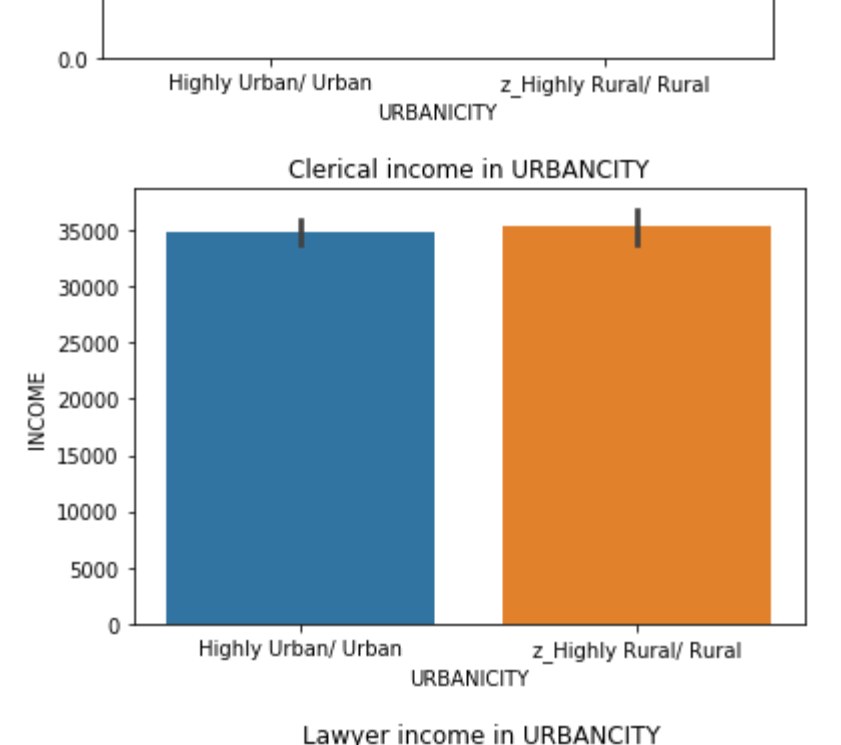
In [345…
```
ax,fig = plt.subplots(figsize=(20,9))
sns.barplot(x=data['CAR_TYPE'],y=data['HOME_VAL'],hue=data['JOB'])
plt.show()
```

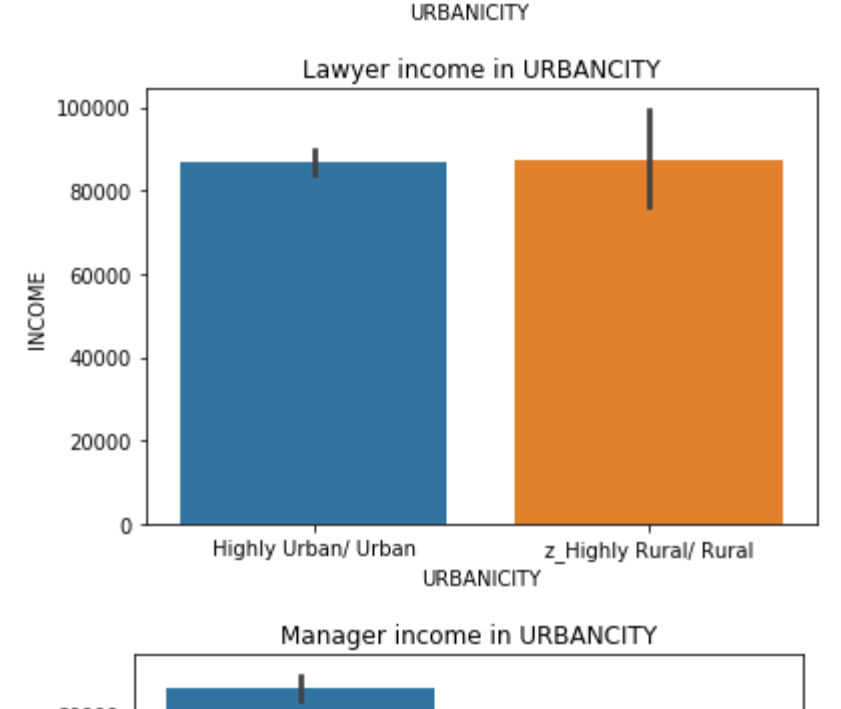


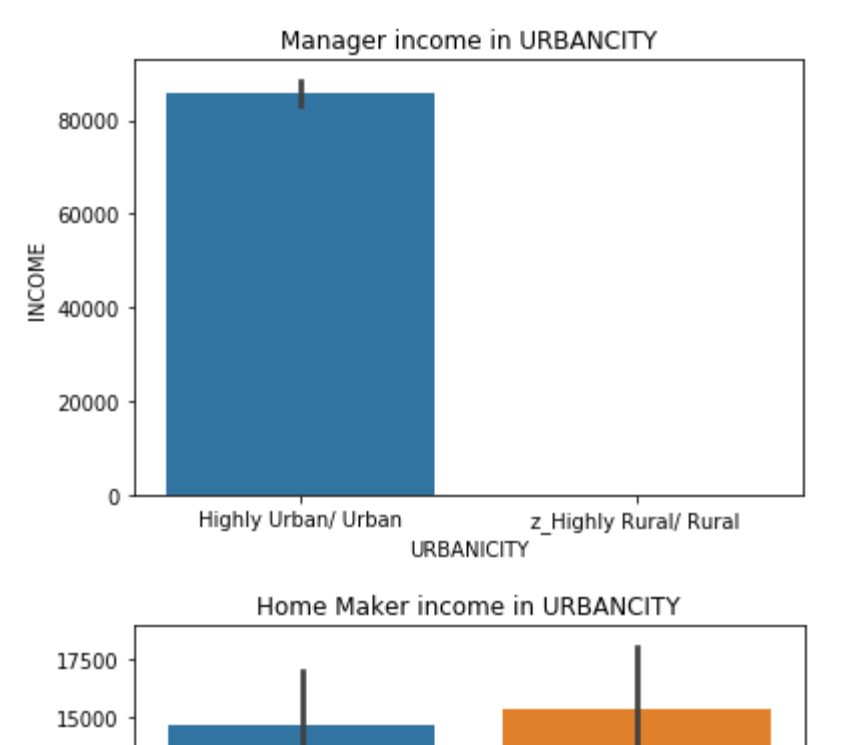
In [346…
```
sns.barplot(data['URBANICITY'],data[data['JOB']=='Doctor']['INCOME'])
plt.title('Doctor income in URBANCITY')
plt.show()
sns.barplot(data['URBANICITY'],data[data['JOB']=='Professional']['INCOME'])
plt.title('Professional income in URBANCITY')
plt.show()
sns.barplot(data['URBANICITY'],data[data['JOB']=='z-Blue-Collar']['INCOME'])
plt.title('z-Blue Collar income in URBANCITY')
plt.show()
sns.barplot(data['URBANICITY'],data[data['JOB']=='Clerical']['INCOME'])
plt.title('Clerical income in URBANCITY')
plt.show()
sns.barplot(data['URBANICITY'],data[data['JOB']=='Lawyer']['INCOME'])
plt.title('Lawyer income in URBANCITY')
plt.show()
sns.barplot(data['URBANICITY'],data[data['JOB']=='Manager']['INCOME'])
plt.title('Manager income in URBANCITY')
plt.show()
sns.barplot(data['URBANICITY'],data[data['JOB']=='Home Maker']['INCOME'])
plt.title('Home Maker income in URBANCITY')
plt.show()
sns.barplot(data['URBANICITY'],data[data['JOB']=='Student']['INCOME'])
plt.title('Student income in URBANCITY')
plt.show()
```

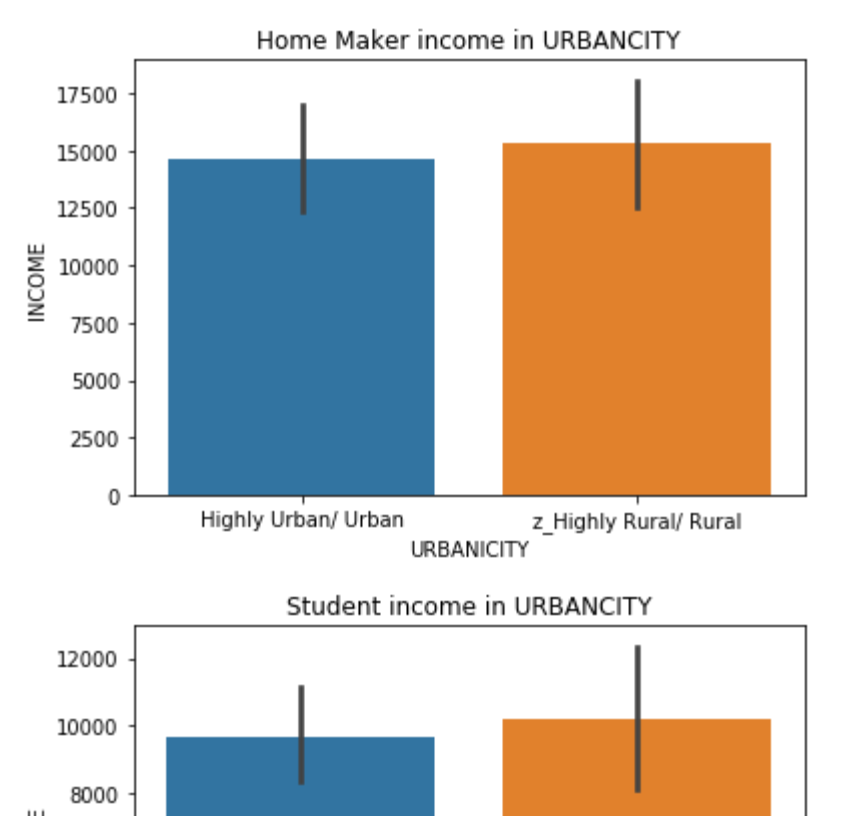


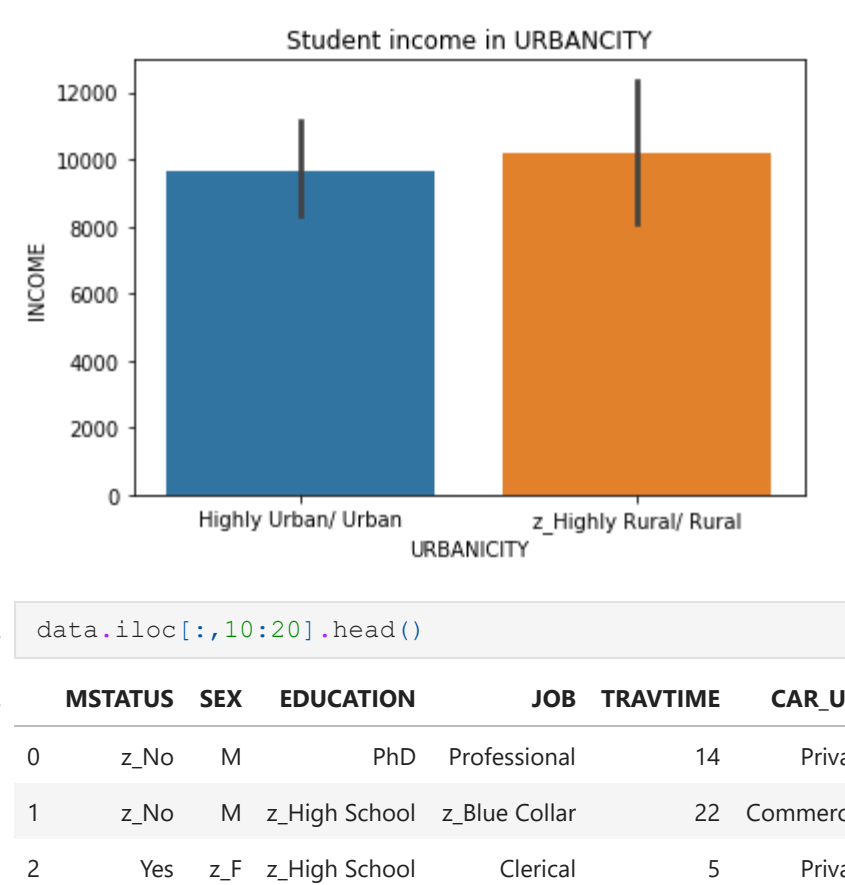


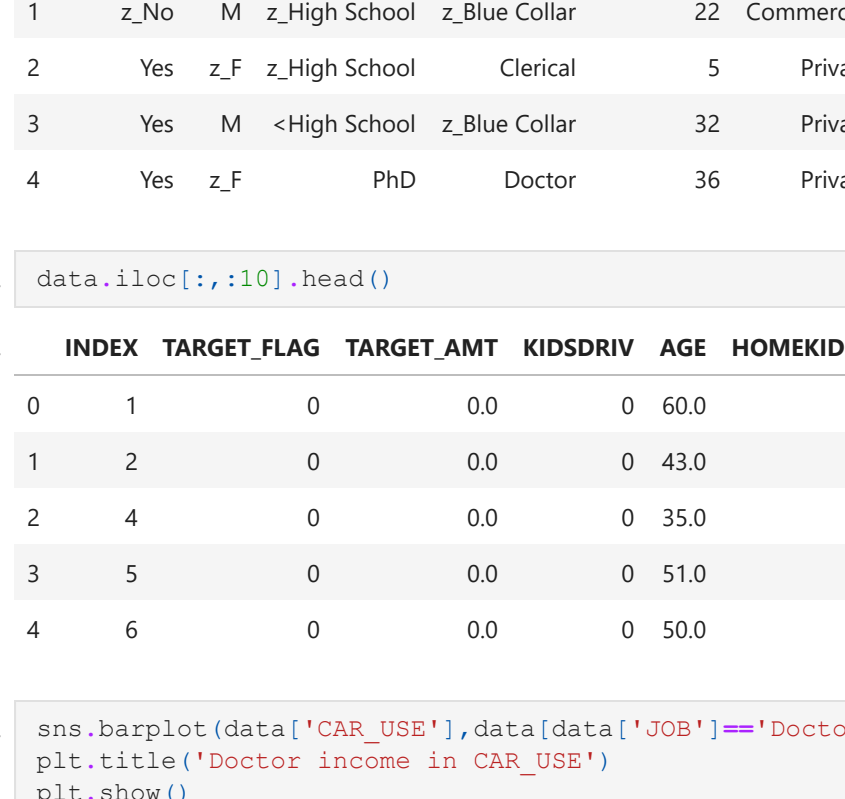












In [347…
```
data.iloc[:,10:20].head()
```

Out[347…

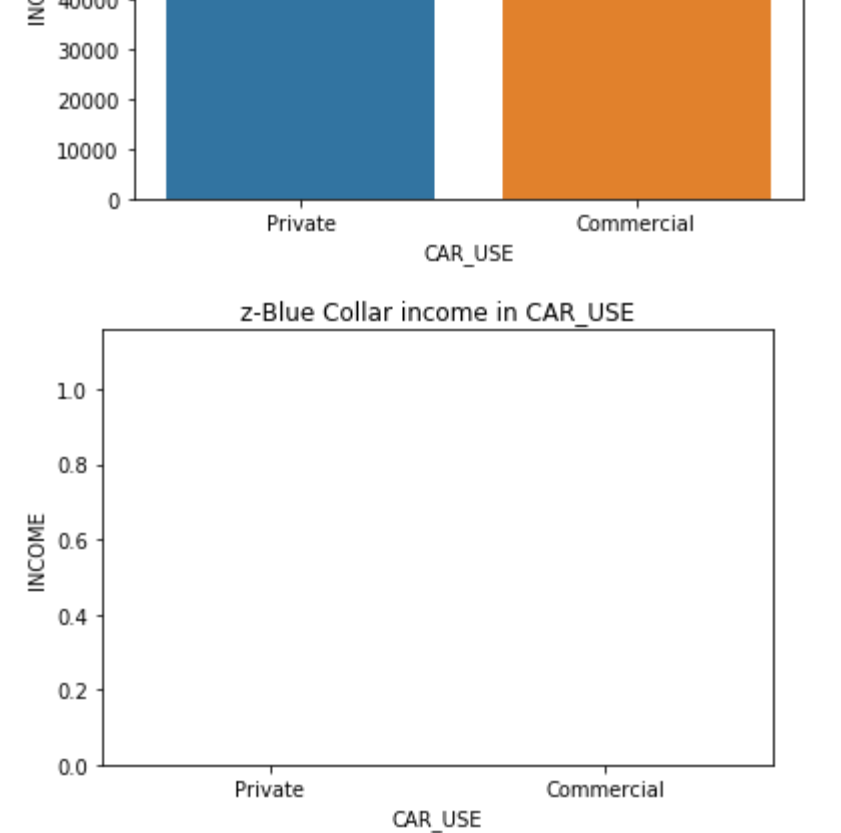
	MSTATUS	SEX	EDUCATION	JOB	TRAVTIME	CAR_USE	BLUEBOOK	TIF	CAR_TYPE	RED_CAR
0	z_No	M	PhD	Professional	14	Private	14230	11	Minivan	yes
1	z_No	M	z_High School	z_Blue Collar	22	Commercial	14940	1	Minivan	yes
2	Yes	z_F	z_High School	Clerical	5	Private	4010	4	z_SUV	no
3	Yes	M	<High School	z_Blue Collar	32	Private	15440	7	Minivan	yes
4	Yes	z_F	PhD	Doctor	36	Private	18000	1	z_SUV	no

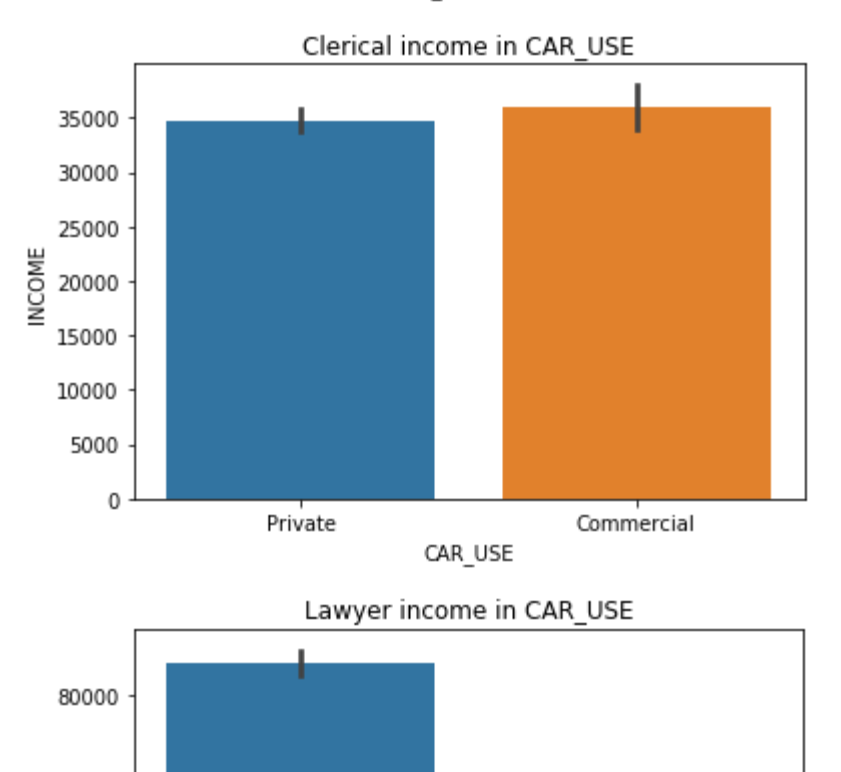
In [348…
```
data.iloc[:,:10].head()
```

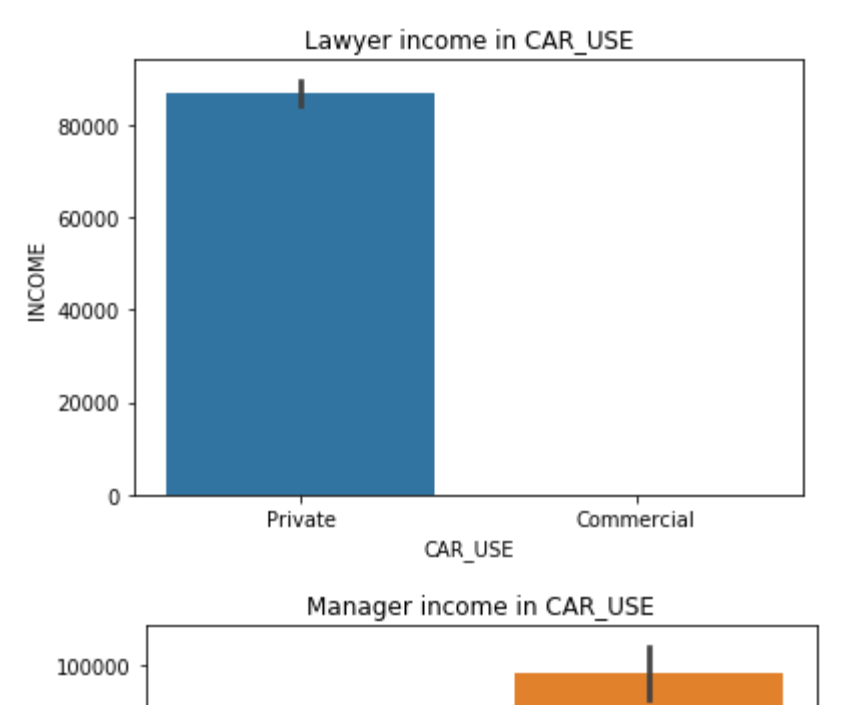
Out[348…

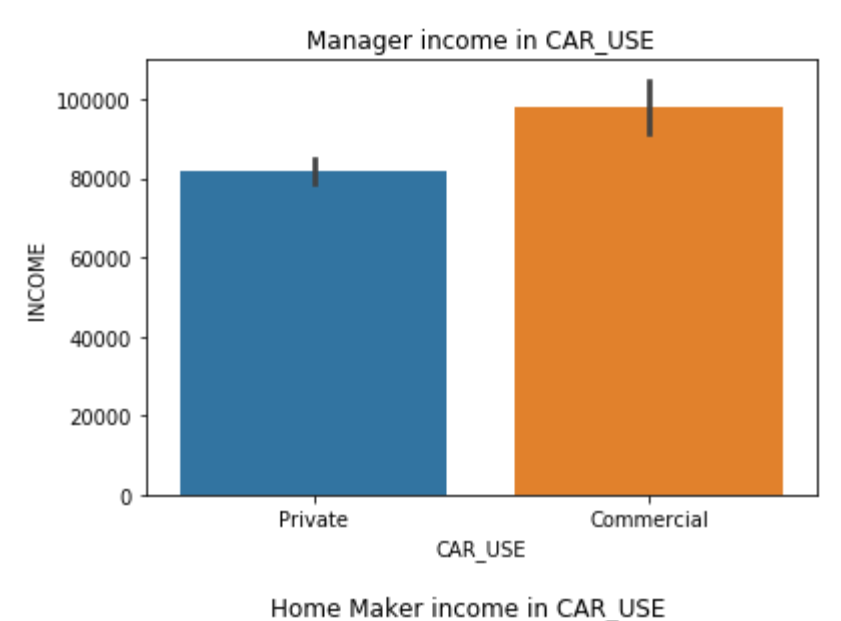
	INDEX	TARGET_FLAG	TARGET_AMT	KIDSDRIV	AGE	HOMEKIDS	YOJ	INCOME	PARENT1	HOME_VAL
0	1	0	0.0	0	60.0	0	11.0	67349.0	No	0.0
1	2	0	0.0	0	43.0	0	11.0	91449.0	No	257252.0
2	4	0	0.0	0	35.0	1	10.0	16039.0	No	124191.0
3	5	0	0.0	0	51.0	0	14.0	54028.0	No	306251.0
4	6	0	0.0	0	50.0	0	11.0	114986.0	No	243925.0

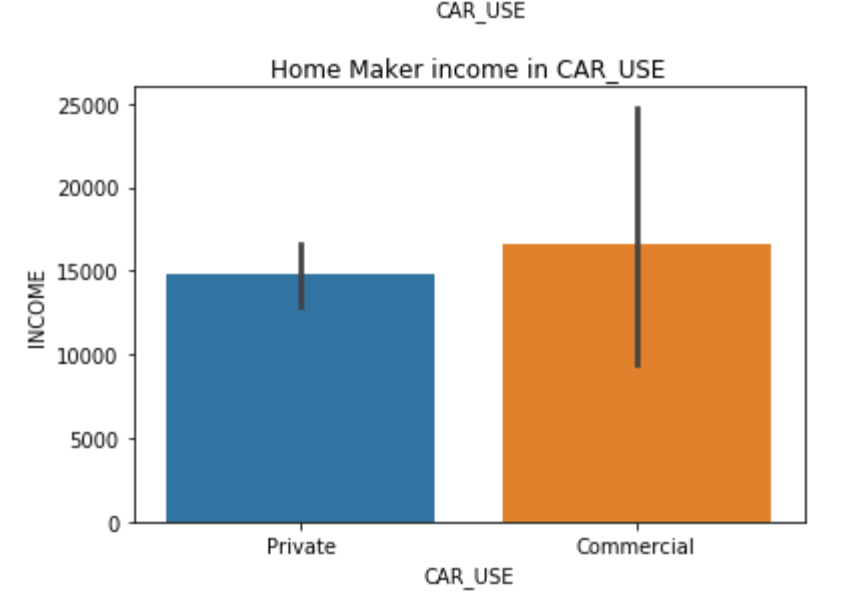
In [349…
```
sns.barplot(data['CAR_USE'],data[data['JOB']=='Doctor']['INCOME'])
plt.title('Doctor income in CAR_USE')
plt.show()
sns.barplot(data['CAR_USE'],data[data['JOB']=='Professional']['INCOME'])
plt.title('Professional income in CAR_USE')
plt.show()
sns.barplot(data['CAR_USE'],data[data['JOB']=='z-Blue-Collar']['INCOME'])
plt.title('z-Blue Collar income in CAR_USE')
plt.show()
sns.barplot(data['CAR_USE'],data[data['JOB']=='Clerical']['INCOME'])
plt.title('Clerical income in CAR_USE')
plt.show()
sns.barplot(data['CAR_USE'],data[data['JOB']=='Lawyer']['INCOME'])
plt.title('Lawyer income in CAR_USE')
plt.show()
sns.barplot(data['CAR_USE'],data[data['JOB']=='Manager']['INCOME'])
plt.title('Manager income in CAR_USE')
plt.show()
sns.barplot(data['CAR_USE'],data[data['JOB']=='Home Maker']['INCOME'])
plt.title('Home Maker income in CAR_USE')
plt.show()
sns.barplot(data['CAR_USE'],data[data['JOB']=='Student']['INCOME'])
plt.title('Student income in CAR_USE')
plt.show()
```

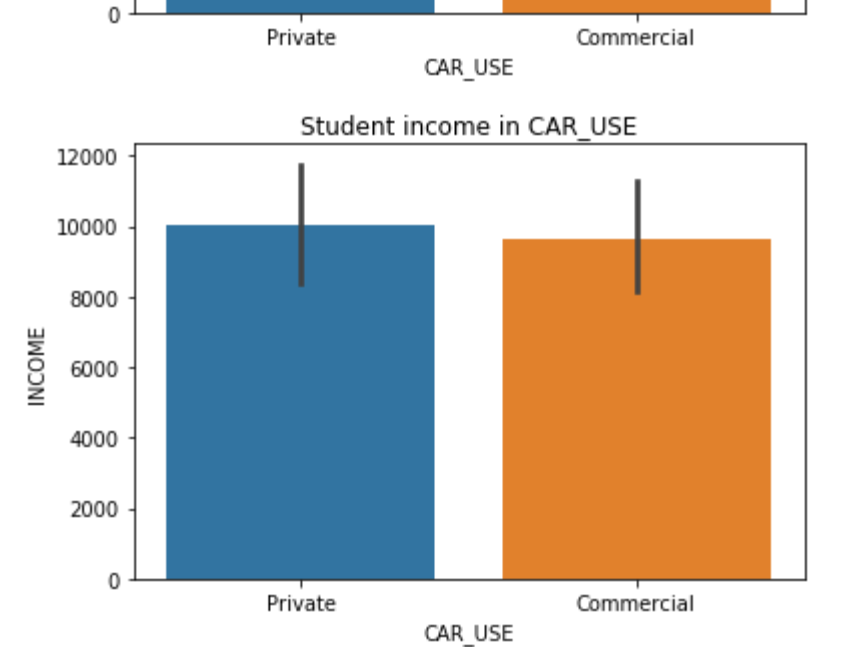


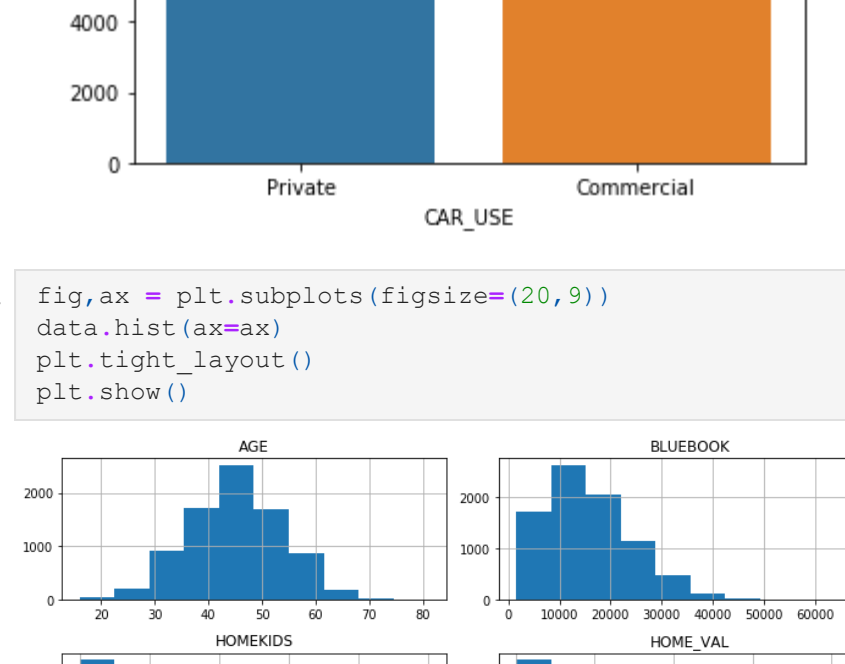




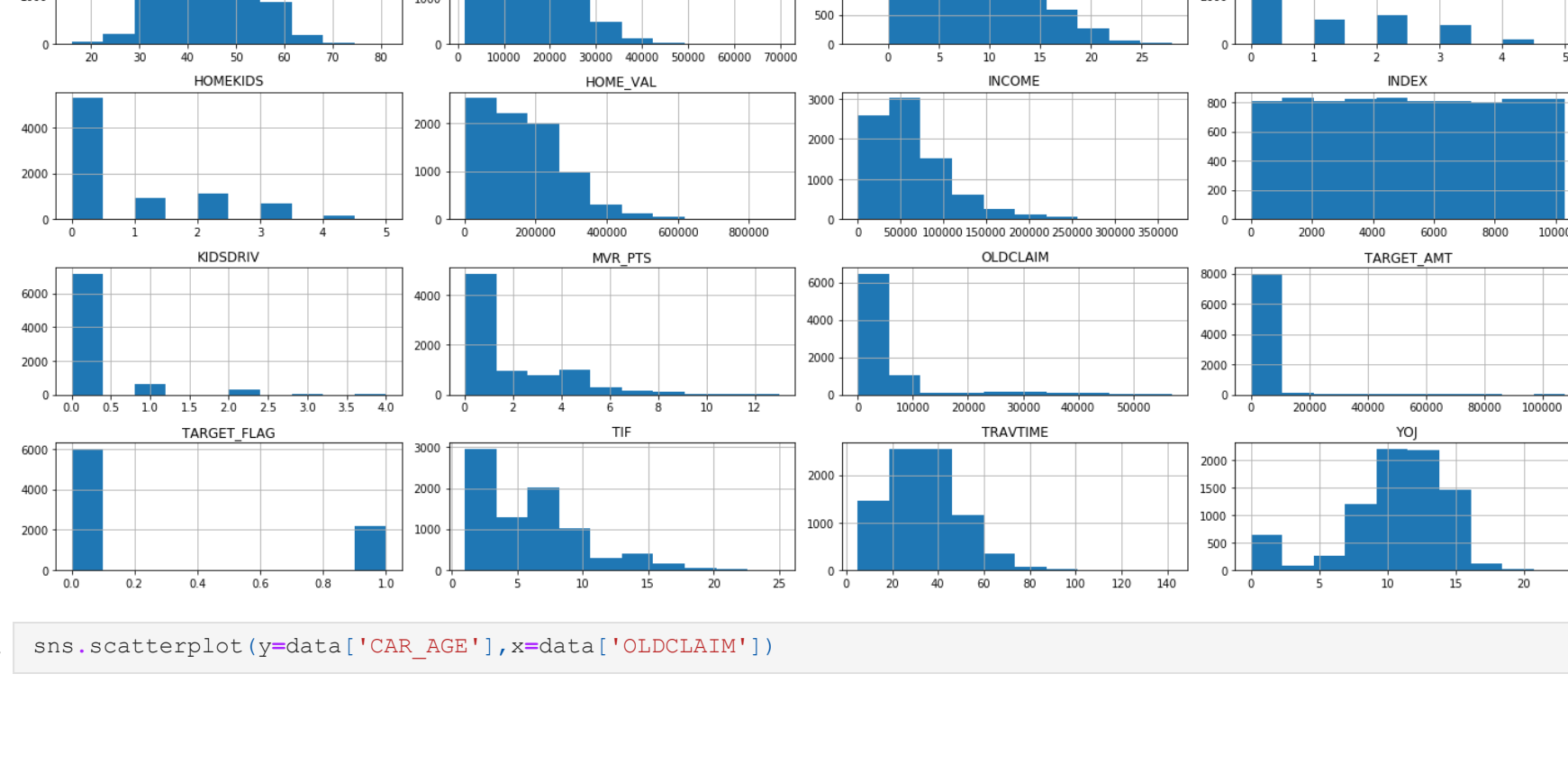






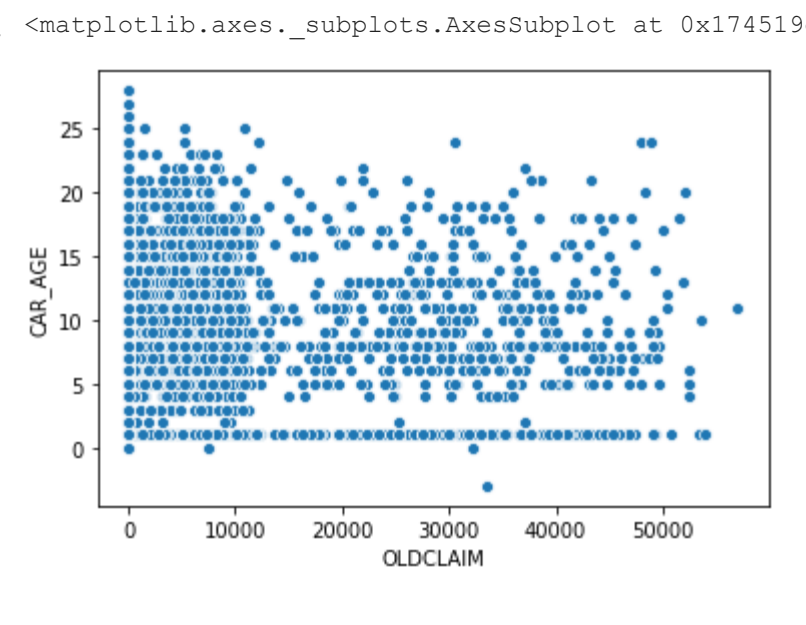


In [350…
```
fig,ax = plt.subplots(figsize=(20,9))
data.hist(ax=ax)
plt.tight_layout()
plt.show()
```



In [351…
```
sns.scatterplot(y=data['CAR_AGE'],x=data['OLDCLAIM'])
```

Out [351… <matplotlib.axes._subplots.AxesSubplot at 0x174519d0>



Data -Preprocessing.

In [352…
```
data.head()
```

Out [352…

	INDEX	TARGET_FLAG	TARGET_AMT	KIDSDRIV	AGE	HOMEKIDS	YOJ	INCOME	PARENT1	HOME_VAL	...	BLUEBOOK	TIF	CAR_TYPE	R
0	1	0	0.0	0	60.0	0	11.0	67349.0	No	0.0	...	14230	11	Minivan	
1	2	0	0.0	0	43.0	0	11.0	91449.0	No	257252.0	...	14940	1	Minivan	
2	4	0	0.0	0	35.0	1	10.0	16039.0	No	124191.0	...	4010	4	z_SUV	
3	5	0	0.0	0	51.0	0	14.0	54028.0	No	306251.0	...	15440	7	Minivan	
4	6	0	0.0	0	50.0	0	11.0	114986.0	No	243925.0	...	18000	1	z_SUV	

5 rows × 26 columns

In [353…
```
from sklearn.preprocessing import LabelEncoder
label = LabelEncoder()
```

In [354…
```
data['PARENT1']= label.fit_transform( data['PARENT1'])
```

In [355…
```
data['CAR_TYPE'].value_counts()
```

Out [355…
```
z_SUV          2294
Minivan        2145
Pickup         1389
Sports Car      907
Van             750
Panel Truck     676
Name: CAR_TYPE, dtype: int64
```

In [356…
```
data['CAR_TYPE'] = data['CAR_TYPE'].apply(lambda x:str(x).replace(' ',''))
data['CAR_TYPE'].value_counts()
```

Out [356…
```
z_SUV         2294
Minivan       2145
Pickup        1389
SportsCar      907
Van            750
PanelTruck     676
Name: CAR_TYPE, dtype: int64
```

In [357…
```
data['CAR_TYPE']= label.fit_transform(data['CAR_TYPE'])
```

In [358…
```
data['RED_CAR'].value_counts()
```

Out [358…
```
no     5783
yes    2378
Name: RED_CAR, dtype: int64
```

In [359…
```
data['RED_CAR']= label.fit_transform(data['RED_CAR'])
```

In [360…
```
data['REVOKED'].value_counts()
```

Out [360…
```
No     7161
Yes    1000
Name: REVOKED, dtype: int64
```

In [361…
```
data['REVOKED']= label.fit_transform(data['REVOKED'])
```

In [362…
```
data['MSTATUS'].value_counts()
```

Out [362…
```
Yes     4894
z_No    3267
Name: MSTATUS, dtype: int64
```

In [363…
```
data['MSTATUS'] = label.fit_transform(data['MSTATUS'])
```

In [364…
```
data['JOB']=label.fit_transform(data['JOB'])
```

In [365…
```
data['CAR_USE']=label.fit_transform(data['CAR_USE'])
```

In [366…
```
data['URBANICITY'].value_counts()
```

Out [366…
```
Highly Urban/ Urban      6492
z_Highly Rural/ Rural    1669
Name: URBANICITY, dtype: int64
```

In [367…
```
data['URBANICITY']= label.fit_transform(data['URBANICITY'])
```

In [368…
```
data['SEX'] = label.fit_transform(data['SEX'])
data['EDUCATION'] = label.fit_transform(data['EDUCATION'])
```

In [369…
```
data.drop(columns=['INDEX'],inplace=True)
```

In [ ]:

In [370…
```
data.iloc[:,10:20]
```

Out [370…

	SEX	EDUCATION	JOB	TRAVTIME	CAR_USE	BLUEBOOK	TIF	CAR_TYPE	RED_CAR	OLDCLAIM
0	0	3	5	14	1	14230	11	0	1	4461
1	0	4	7	22	0	14940	1	0	1	0
2	1	4	0	5	1	4010	4	5	0	38690
3	0	0	7	32	1	15440	7	0	1	0
4	1	3	1	36	1	18000	1	5	0	19217
...	...	...	...	...	...	...	...	...	...	...
8156	0	4	7	51	0	27330	10	1	1	0
8157	0	3	4	21	1	13270	15	0	0	0
8158	0	2	7	36	0	24490	6	1	0	0
8159	1	1	2	36	1	22550	6	0	0	0
8160	1	4	0	64	1	19400	6	0	0	0

8161 rows × 10 columns

In [371…
```
data['CAR_AGE'].mean()
```

Out [371… 8.30780541600294

In [372…
```
data['CAR_AGE'].median()
```

Out [372… 8.0

In [373…
```
sns.distplot(data['CAR_AGE'])
```

Out [373… <matplotlib.axes._subplots.AxesSubplot at 0x170a7590>

In [374…
```
data['CAR_AGE'].describe()
```

Out [374…
```
count    8161.000000
mean        8.307805
std         5.520292
min        -3.000000
25%         4.000000
50%         8.000000
75%        12.000000
max        28.000000
Name: CAR_AGE, dtype: float64
```

In [375…
```
category = []

col = data['CAR_AGE']

for i in col:
    if i >9:
        category.append('No')
    else:
        category.append('Yes')
```

In [376…
```
category
```

```
 'No',
 'Yes',
 'Yes',
 'Yes',
 'Yes',
 'Yes',
 'Yes',
 'No',
 'Yes',
 'No',
 'Yes',
 'No',
 'Yes',
 'Yes',
 'Yes',
 'Yes',
 'No',
 'No',
 'No',
 'Yes',
 'Yes',
 'Yes',
 'No',
 'No',
 'Yes',
 'Yes',
 'Yes',
 'Yes',
 'Yes',
 'Yes',
 'Yes',
 'No',
 'No',
 'No',
 'No',
 'Yes',
 'Yes',
 'No',
 'No',
 'No',
 'No',
 ...]
```

In [377…
```
data['Category'] = category
```

In [378…
```
data['Category'].value_counts()
```

Out[378…
```
Yes    4992
No     3169
Name: Category, dtype: int64
```

In [379…
```
data['Category'] = label.fit_transform(data['Category'])
```

In [380…
```
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score,classification_report,confusion_matrix
from sklearn.preprocessing import StandardScaler
from sklearn.feature_selection import chi2
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.feature_selection import VarianceThreshold
from sklearn.model_selection import RandomizedSearchCV
```

In [381…
```
x = data.drop(columns='Category')
y = data['Category']
```

In [382…
```
x.shape,y.shape
```

Out[382…
```
((8161, 25), (8161,))
```

In [383…
```
x_train,x_test,y_train,y_test = train_test_split(x,y,train_size=0.80,stratify =y,random_state=42)
```

In [384…
```
x_train.shape,x_test.shape
```

Out[384…
```
((6528, 25), (1633, 25))
```

In [385…
```
y_train.value_counts()
```

Out[385…
```
1    3993
0    2535
Name: Category, dtype: int64
```

In [412…
```
thresh = VarianceThreshold(threshold=0.1)
x_train_unique = thresh.fit_transform(x_train)
x_test_unique  = thresh.transform(x_test)
```

In [413…
```
x_train_unique.shape,x_test_unique.shape
```

Out[413…
```
((6528, 25), (1633, 25))
```

In [414…
```
x_train_T = x_train_unique.T
x_test_T   = x_test_unique.T
```

In [415…
```
x_train_T = pd.DataFrame(x_train_T)
x_test_T  = pd.DataFrame(x_test_T)
```

In [416…
```
duplicated = x_train_T.duplicated().sum()
duplicated
```

Out[416…
```
0
```

In [417…
```
from imblearn.over_sampling import SMOTE
```

In [418…
```
sm = SMOTE(sampling_strategy=1,random_state=42,k_neighbors=5)
x_train_res,y_train_res = sm.fit_sample(x_train_unique,y_train)
```

In [419…
```
x_train_res.shape,y_train_res.shape
```

Out[419…
```
((7986, 25), (7986,))
```

In [420…
```
print('After the SMOTE')
print('y_train_res with 0',len(y_train_res[y_train_res==0]))
print('y_train_res with 1',len(y_train_res[y_train_res==1]))
```

```
After the SMOTE
y_train_res with 0 3993
y_train_res with 1 3993
```

Model Selection

In [421…
```
model = RandomForestClassifier(n_estimators=100,random_state=42)
model.fit(x_train_res,y_train_res)
y_pred = model.predict(x_test_res)
print('Accuracy :',accuracy_score(y_test,y_pred))
```

```
Accuracy : 1.0
```

In [ ]:

In [ ]: