ACTU PS5841 Data Science in Finance & Insurance - Autumn 2022 (Y. Wang)
Assignment - 5
Assigned 9/29/22, Due 10/8/22

The file `Carseats.csv` records child car seat sales in 400 locations. The following linear regression model attempts to predict `Sales` in non-US locations (`US = No`):

`Sales` ~ `Income + Price + ShelveLoc + Urban + Urban:Income`

where the categorical feature `ShelveLoc` is coded according to the sum-to-zero contrast, and `Urban` is coded according to the treatment contrast.

We can easily fit the regression model in python using `statsmodels` as follows.

```
from patsy.contrasts import Treatment, Sum
import statsmodels.formula.api as smf   #smf.ols
```

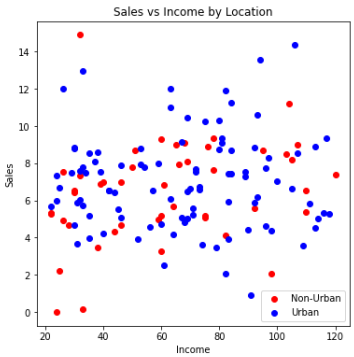
```
sum_contrast = Sum().code_without_intercept(['Bad', 'Good', 'Medium'])
treatment_contrast = Treatment(reference = 'No').code_without_intercept(['No', 'Yes'])
lm_smf_res = smf.ols("Sales ~ Income + Price +\
               C( ShelveLoc, sum_contrast) + C(Urban, treatment_contrast) +\
               C(Urban, treatment_contrast):Income",\
              data = df).fit()
lm_smf_res.summary()
```

Write a python script to "manually" validate the fit. You can only use `numpy`, `pandas`, and `matplotlib`. Please let your code output the following.
[*a*] The sales per non-US geographic location by Shelve Location (`ShelveLoc`) and Location Type (`Urban`) in a table:

Average Sales

ShelveLoc	Urban	NonUrban
Bad	5.359130	5.135455
Medium	6.958909	6.440690
Good	9.210556	8.968333

[*b*] A scatter plot of `Sales` vs `Income` by `Urban`



Page 2 is the last page.

[c] The regression coefficient estimates as well as R^2:

	coef
Income	0.026715
Price	-0.057761
ShelveLoc_Bad	-1.956725
ShelveLoc_Good	2.297505
Urban_Yes	1.887842
Urban:Income	-0.024027
Intercept	11.757081

```
R-squared = 0.577065
```

Please submit your work as hw5.ipynb and hw5.html to `Canvas`.