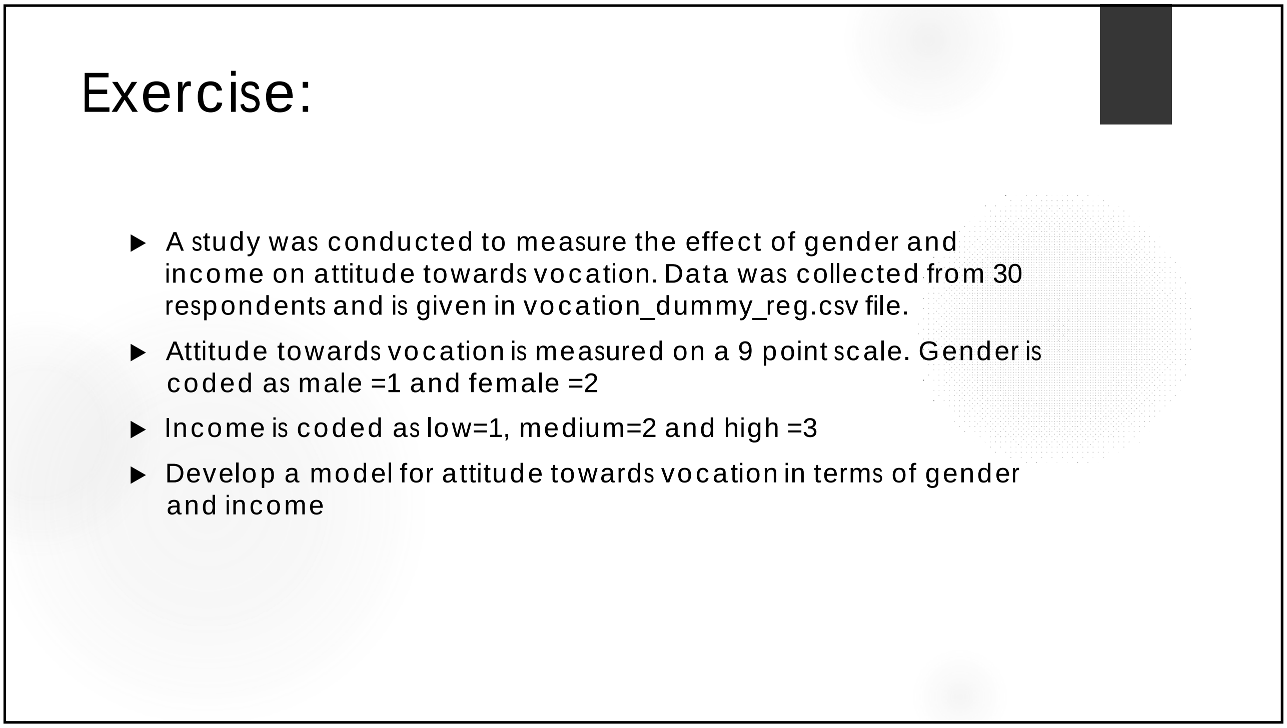
REGRESSION WITH DUMMY VARIABLES

MOHAN M J

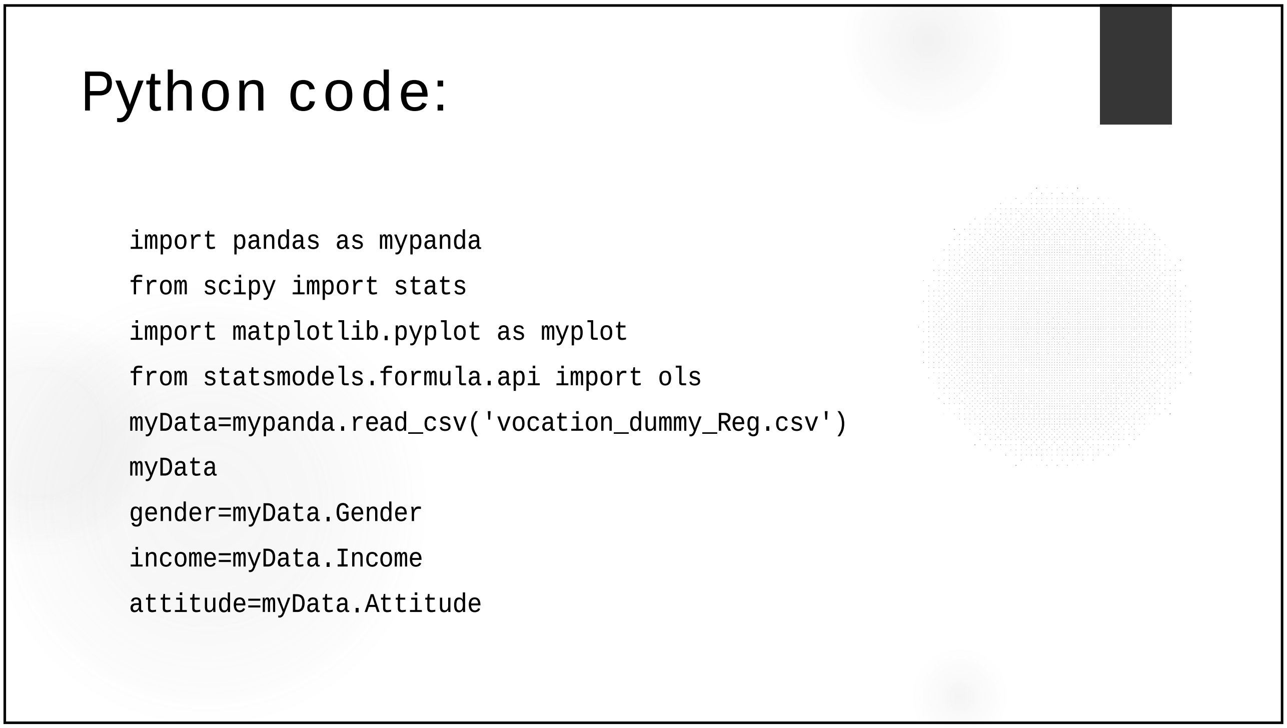
Introduction

- When Xs are not numeric but nominal
- Each nominal or categorical variable is converted into dummy variables
- Dummy Variables will take values 0 or 1
- Number of dummy variables for one X variable is equal to number of distinct values of that variable - 1



Exercise:

- A study was conducted to measure the effect of gender and income on attitude towards vocation. Data was collected from 30 respondents and is given in vocation_dummy_reg.csv file.
- Attitude towards vocation is measured on a 9 point scale. Gender is coded as male =1 and female =2
- Income is coded as low=1, medium=2 and high =3
- Develop a model for attitude towards vocation in terms of gender and income

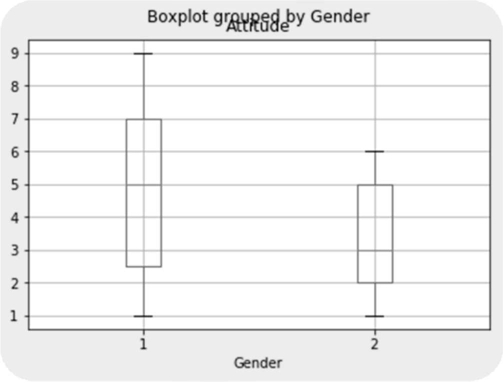


Python code:

```
import pandas as mypanda
from scipy import stats
import matplotlib.pyplot as myplot
from statsmodels.formula.api import ols
myData=mypanda.read_csv('vocation_dummy_Reg.csv')
myData
gender=myData.Gender
income=myData.Income
attitude=myData.Attitude
```

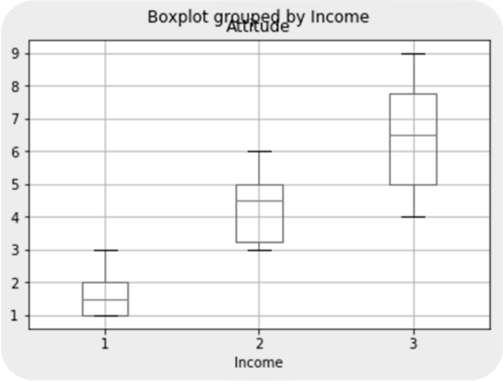
Python code:

```
myData.boxplot(column='Attitude', by='Gender')
myplot.show()
myData.boxplot(column='Attitude', by='Income')
myplot.show()
```



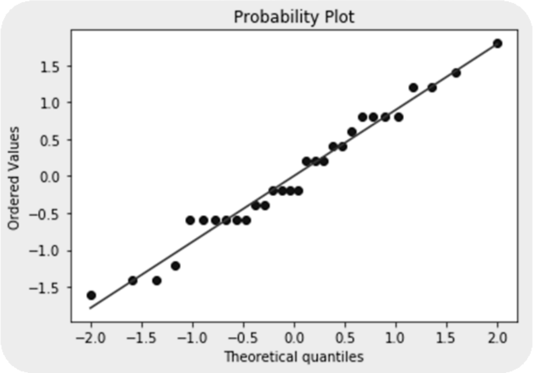
Python code:

```
myData.boxplot(column='Attitude', by='Income')
myplot.show()
```



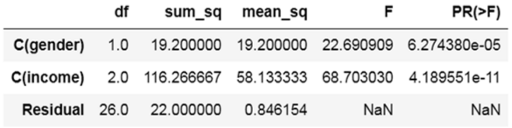
Python code:

```
mymodel=ols('attitude ~ C(gender)+C(income)',myData).fit()
mymodel.summary()
pred=mymodel.predict()
pred
res=attitude-pred
stats.probplot(res,plot=myplot)
myplot.show()
```



Python code:

```
stats.normaltest(res)
Out[] NormaltestResult(statistic=0.52111989611555032,
   pvalue=0.7706199578215539)
from statsmodels.stats.anova import anova_lm
anova_table = anova_lm(mymodel)
anova_table
```



	df	sum_sq	mean_sq	F	PR(>F)
C(gender)	1.0	19.200000	19.200000	22.690909	6.274380e-05
C(income)	2.0	116.266667	58.133333	68.703030	4.189551e-11
Residual	26.0	22.000000	0.846154	NaN	NaN

THANKS