DATA VISUALISATION

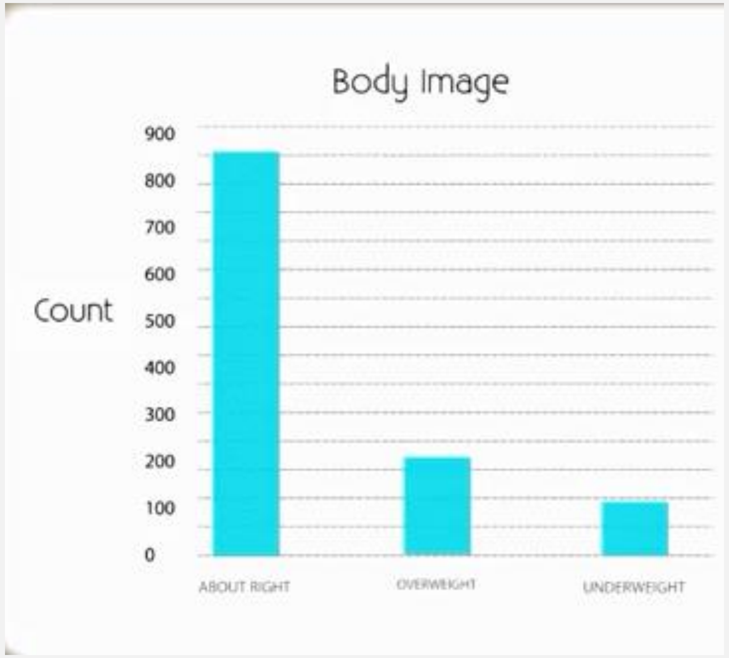
DATA VISUALISATION

- Descriptive statistics
- Univariate graphing
- Bivariate graphing
- Lab7

DATA VISUALISATION

- To visualise the variables we use graphs. Starting with graphing one variable at a time.
- Bar charts are the most commonly used graph to examine the distribution of individual variables.
- The x axis is where we plot the variable we are visualising. In this first case it is a categorical variable. The Y axis is therefore the count of the number of occurrences for each possible value in the categorical variable.

DATA VISUALISATION



DATA VISUALISATION

- To visualise in Python we must import additional libraries into our program.
- First the seaborn package and secondly the matplotlib.pyplot library.
- To ensure a variable is ordered properly on the X axis (horizontal), you should always convert your categorical variables, which are often formatted as numeric variables. We convert them into a category variable so that Python recognises them.

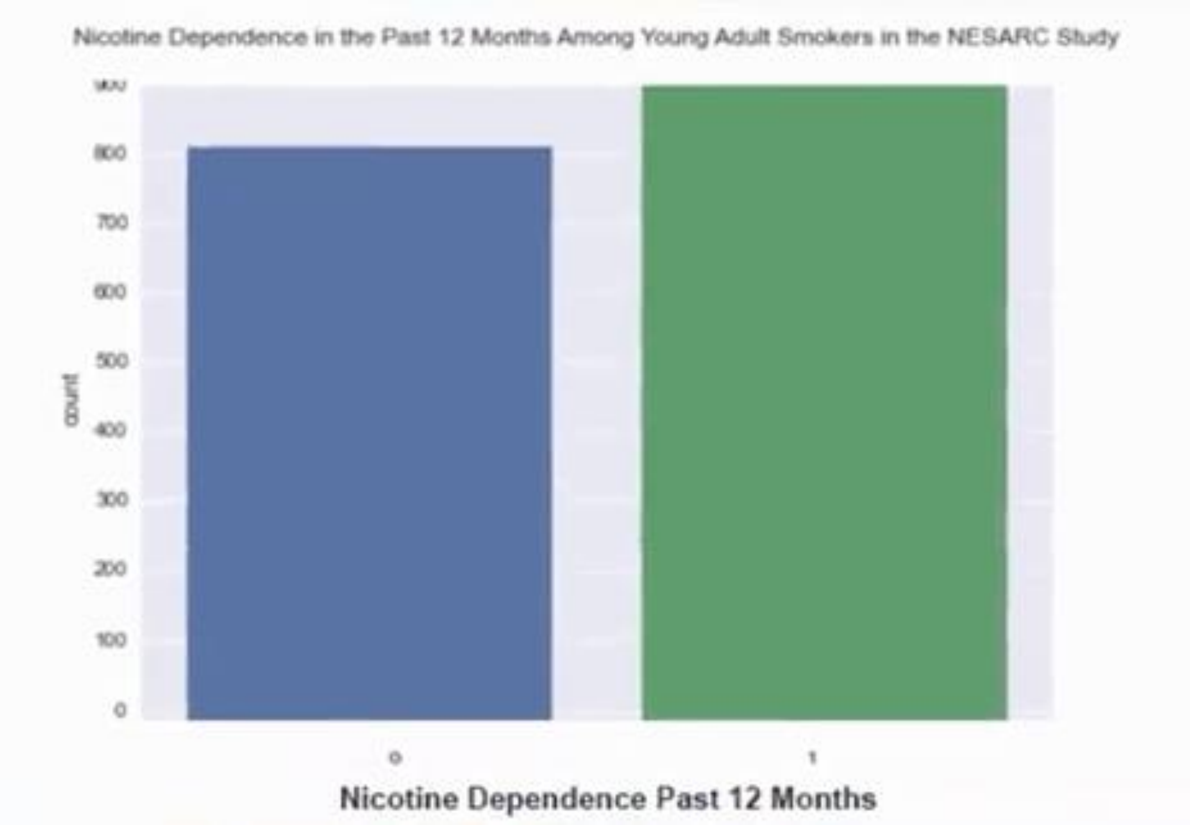
```
#change format of variable from numerical to
Categorical

subset2[‘TAB12MDX'] =
subset2[‘TAB12MDX’].astype('category')
```

DATA VISUALISATION

```
bc1 = seaborn.countplot(x='TAB12MDX',data=subset2)
plt.xlabel('Nicotine Dependence past 12 months')
plt.title('Nicotine Dependence in the past 12 months
among young adult smokers in the Nesarc study')
```

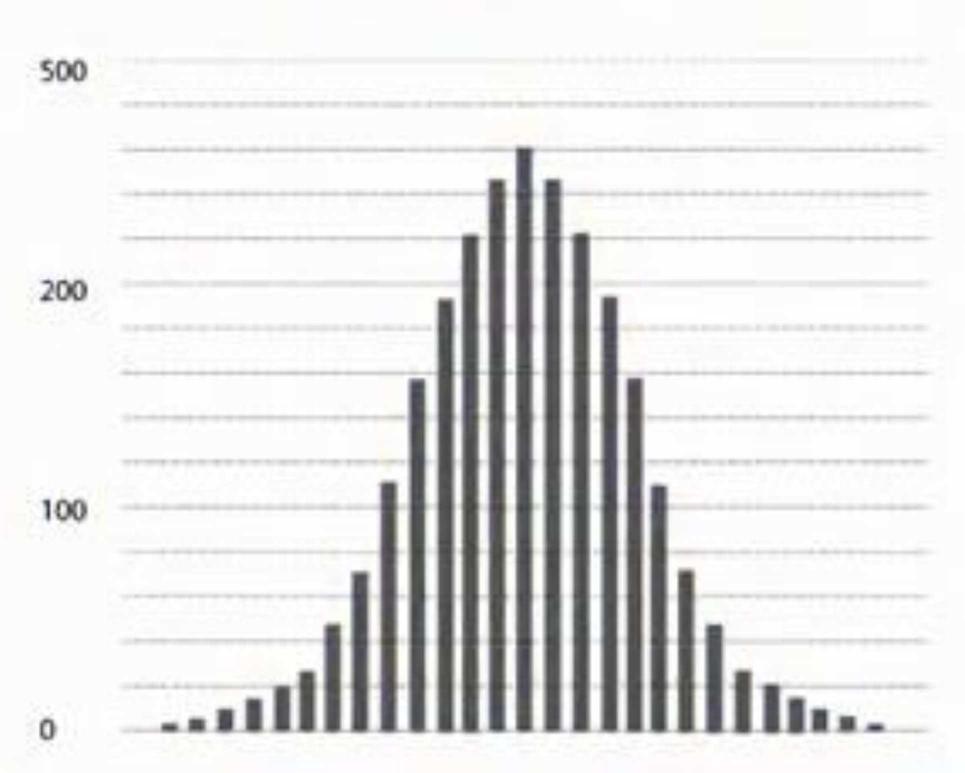
DATA VISUALISATION



DATA VISUALISATION

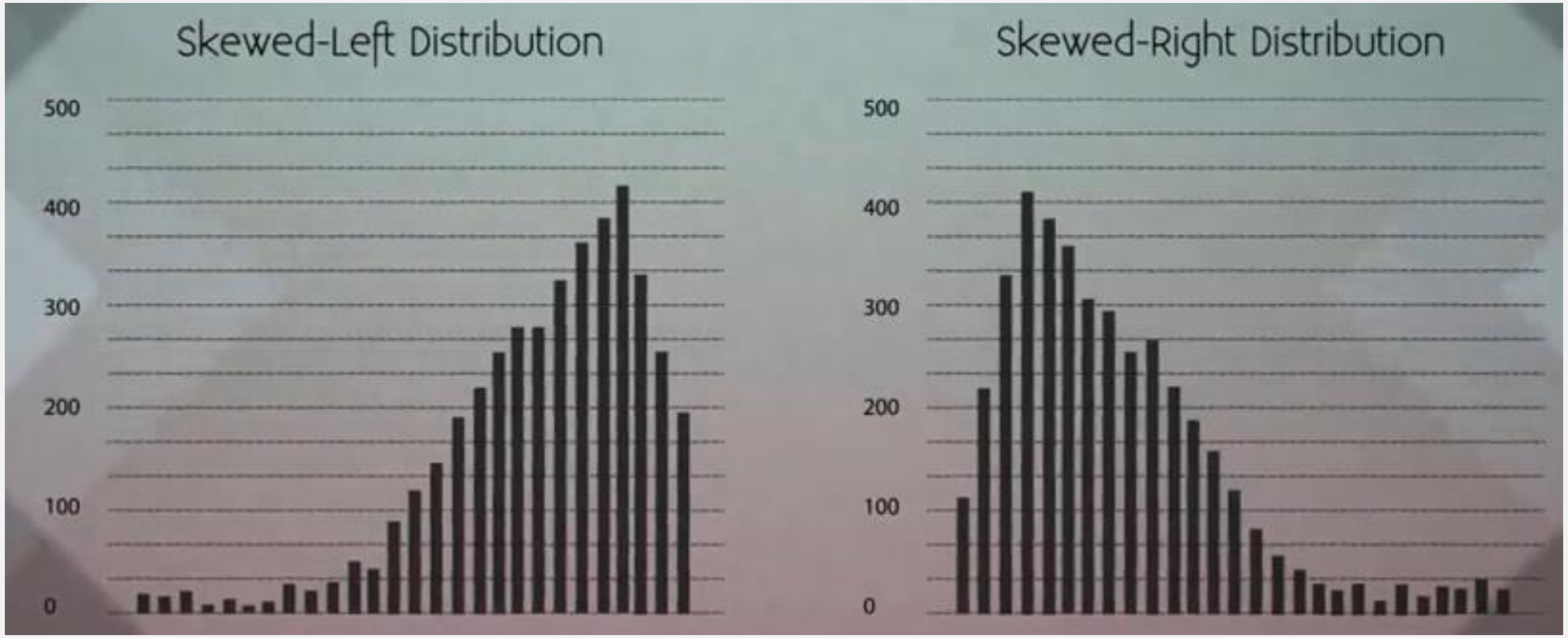
- Some numerical variables are best described and visualised by creating groups for the variable and thus converting it into a categorical variable.
- For exam student grades could be naturally grouped into grand bands and the distribution of the grade bands visualised.
- The X axis would be the grade bands (categorical) and the Y axis would be the number of observations for each grade band.
- Once visualised you can interpret the histogram based on your knowledge from last term.

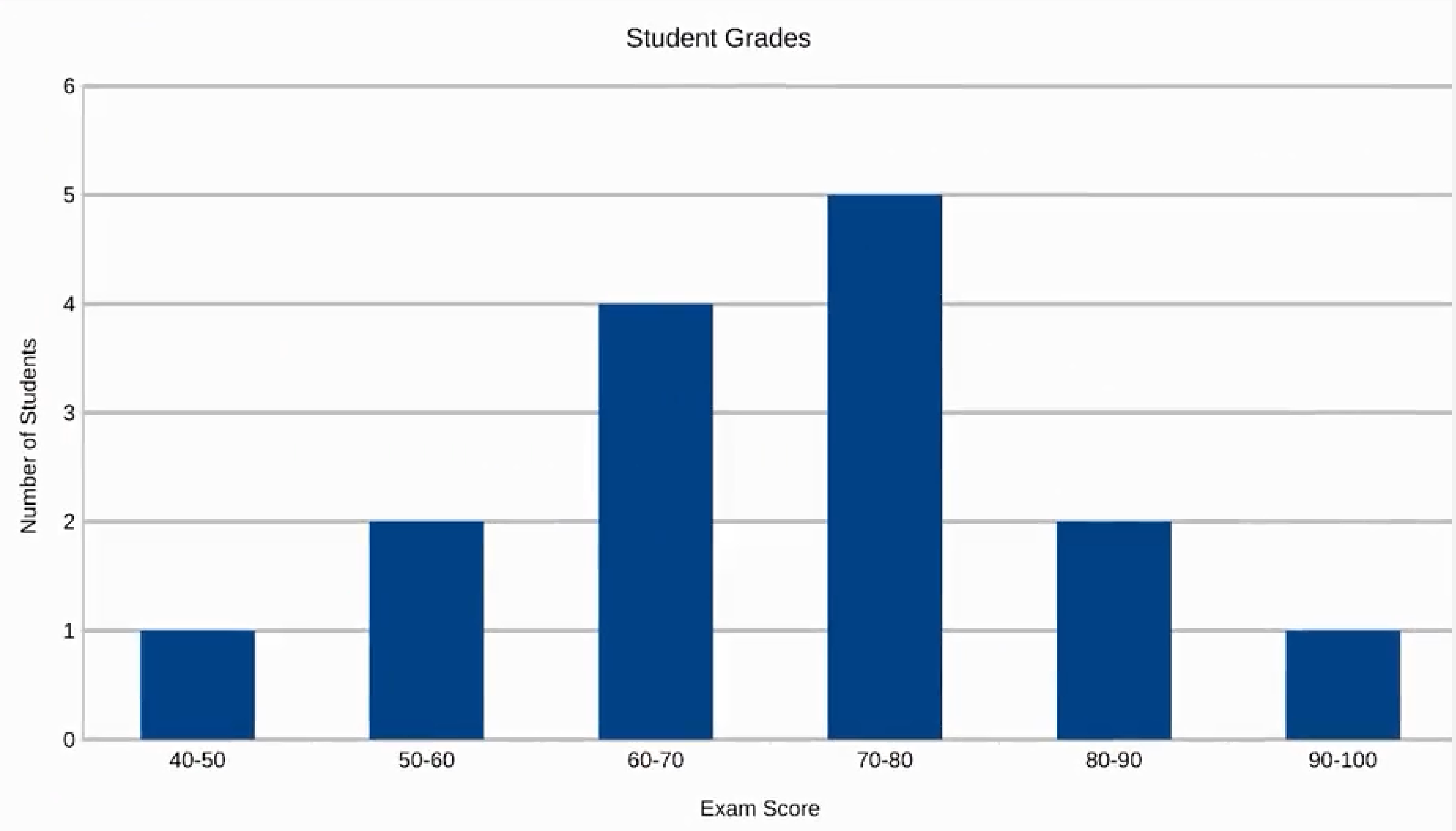
DATA VISUALISATION



What can you say about this histogram?

DATA VISUALISATION




Student Grades
Number of Students
0
1
2
3
4
5
6
40-50
50-60
60-70
70-80
80-90
90-100
Exam Score

DATA VISUALISATION

- It is important to supplement a graphical display with some measures of spread, shape, and variability.

DATA VISUALISATION

```
print('describe number of cigarettes smoked per month')
desc1 = subset2['NUMCIGMO_EST'].describe()
print(desc1)


print('mean')
mean1 = subset2['NUMCIGMO_EST'].mean()
print(mean1)


print('std deviation')
std1 = subset2['NUMCIGMO_EST'].std()
print(std1)
```

```
print('min')
min1 = subset2['NUMCIGMO_EST'].min()
print(min1)

print('max')
max1 = subset2['NUMCIGMO_EST'].max()
print(max1)

print('median')
median1 = subset2['NUMCIGMO_EST'].median()
print(median1)

print('mode')
mode1 = subset2['NUMCIGMO_EST'].mode()
print(mode1)
```

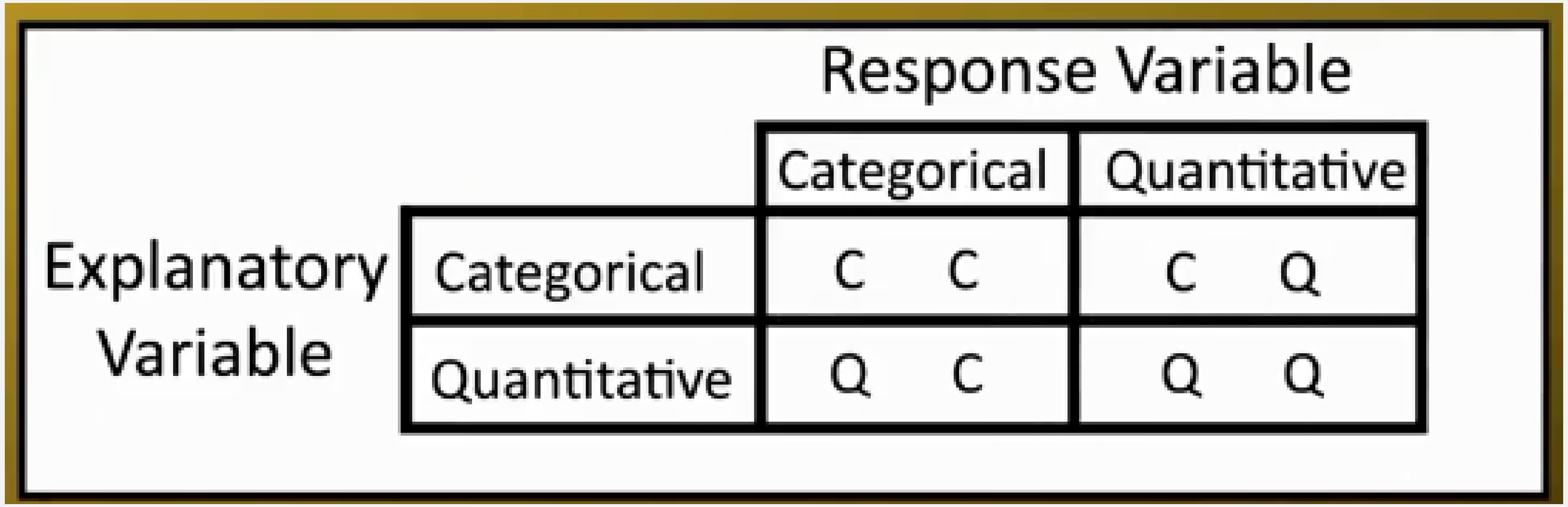
DATA VISUALISATION

- The describe function works well on a numerical variable, calculating the mean, mode etc.
- What about using the describe function on a categorical variable?
- It will calculate a count, the number of unique values, the top or highest value and the frequency of that top value.
- You must remember to use the appropriate descriptive statistics for both quantitative and categorical variables.
- For quantitative it is best to examine histograms, and then supplement with measures of shape, centre and spread.
- Categorical variables often described well with frequency distributions or with a bar chart.

DATA VISUALISATION

- Visualising the relationship between two variables:
- Each variable has a role to play.
- A variable may either be a response variable, also known as the dependent variable or outcome variable, or it could be the explanatory variable, also known as the independent variable or predictor variable.
- The following classification helps in selecting statistical tolls that can be used to explore the relationship of variables and selecting the appropriate graphs.

DATA VISUALISATION



		Response Variable	
		Categorical	Quantitative
Explanatory Variable	Categorical	C C	C Q
	Quantitative	Q C	Q Q

DATA VISUALISATION

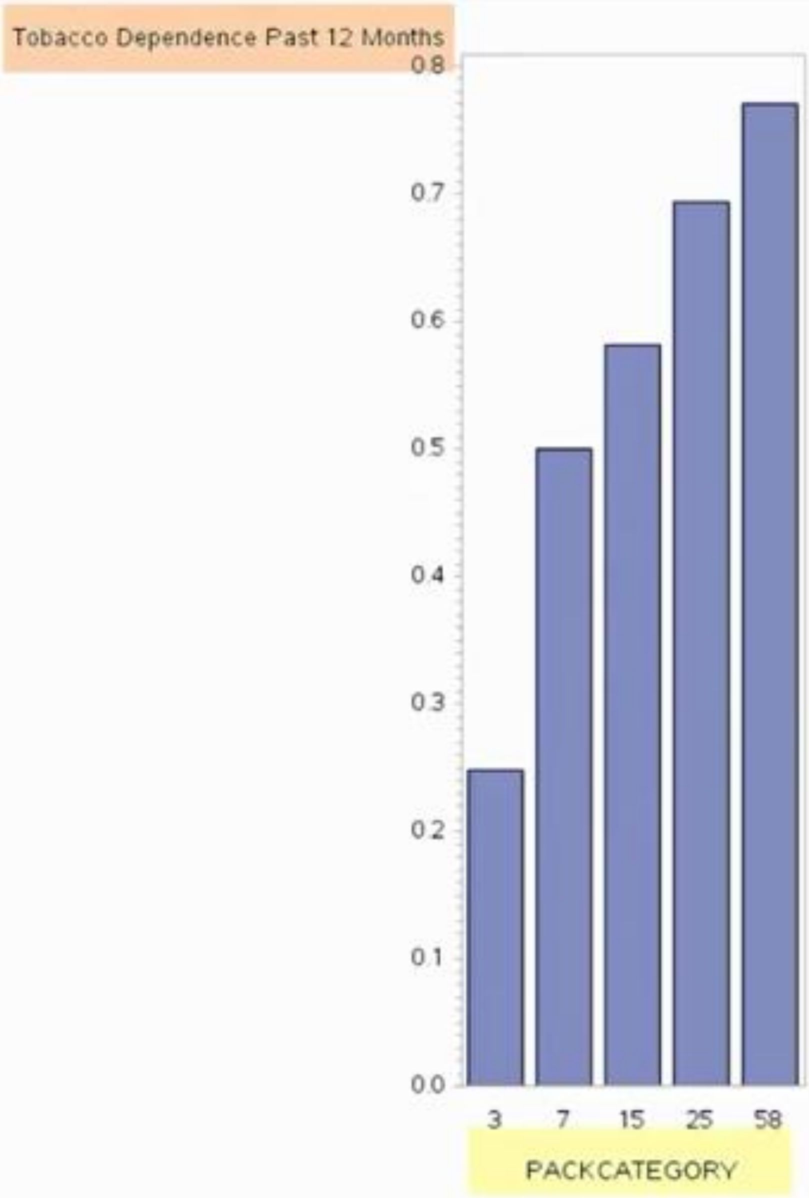
- Explanatory variable goes on the X axis and the response variable goes on the Y axis (usually count/average for each explanatory variable)
- For our example smoking will be the explanatory variable and nicotine dependence the response variable.
- To graph correctly we must first answer a couple of questions.
- What type is the response variable? In our case it is categorical
- How many categories are in the response variable? 1 and 0 (yes and no)
- What type is the explanatory variable? Number of cigarettes smoked per month, so it is a quantitative variable.

DATA VISUALISATION

- When you have a response variable that is a categorical and an explanatory variable that is quantitative then we need to convert the quantitative variable into categorical.
- In our case that means grouping the number of cigarettes smoked per month.
- This would mean we could then graph a categorical by categorical chart.
- The factorplot() is used to graph two categorical variables.

DATA VISUALISATION

- For a bivariate graph where the response variable (TAB12MDX) is categorical with two values 0 and 1 we convert it to numerical so that the Y Axis shows the mean for each observation that falls into the explanatory variable categories. Otherwise Python won't calculate the mean if it is a categorical variable.
- For example, the number of 1's (TAB12MDX) are added together for all observations that fall into the packcategory value of 1-5 packs per month and divided by the total number of observations, thus giving the proportion of nicotine dependent young adults who smoke 1-5 packs per month.


Tobacco Dependence Past 12 Months
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
3
7
15
25
58
PACKCATEGORY