Collecting the data:

The dataset for 5000 customers, and their 130 attributes (both categorical and numerical) was collected, along with the data dictionary.

Importing libraries:

The necessary libraries, functions and methods were imported.

Reading the data:

Using the read_excel function from Pandas, the data was read into a dataframe object ‘df’ in the notebook.

```
In [4]: df.head()
```

Out[4]:

	custid	region	townsize	gender	age	agecat	birthmonth	ed	edcat	jobcat	...	owncd	ownpda	ownpc
0	3964-QJWTRG-NPN	1	2.0	1	20	2	September	15	3	1	...	0	0	0
1	0648-AIPJSP-UVM	5	5.0	0	22	2	May	17	4	2	...	1	1	1
2	5195-TLUDJE-HVO	3	4.0	1	67	6	June	14	2	2	...	1	0	0
3	4459-VLPQUH-3OL	4	3.0	0	23	2	May	16	3	2	...	1	0	1

Understanding the data:

```
In [10]: df.shape
Out[10]: (5000, 130)

In [11]: df.columns
Out[11]: Index(['custid', 'region', 'townsize', 'gender', 'age', 'agecat', 'birthmonth',
                'ed', 'edcat', 'jobcat',
                ...
                'owncd', 'ownpda', 'ownpc', 'ownipod', 'owngame', 'ownfax', 'news',
                'response_01', 'response_02', 'response_03'],
               dtype='object', length=130)

In [12]: df.info()
         <class 'pandas.core.frame.DataFrame'>
         RangeIndex: 5000 entries, 0 to 4999
         Columns: 130 entries, custid to response_03
```

The data has 5,000 instances and 130 features. There are 31 float features, 97 integer features, and 2 object features.

```
In [13]: df.describe()
```

Out[13]:

	region	townsize	gender	age	agecat	ed	edcat
count	5000.00000	4998.000000	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000
mean	3.00140	2.687275	0.503600	47.025600	4.238800	14.543000	2.672000
std	1.42176	1.425925	0.500037	17.770338	1.308785	3.281083	1.211738
min	1.00000	1.000000	0.000000	18.000000	2.000000	6.000000	1.000000
25%	2.00000	1.000000	0.000000	31.000000	3.000000	12.000000	2.000000
50%	3.00000	3.000000	1.000000	47.000000	4.000000	14.000000	2.000000
75%	4.00000	4.000000	1.000000	62.000000	5.000000	17.000000	4.000000
max	5.00000	5.000000	1.000000	79.000000	6.000000	23.000000	5.000000

8 rows × 128 columns

```
In [14]: df.dtypes
```

Out[14]:

```
custid          object
region           int64
townsize       float64
gender           int64
```

The describe function only gives a summary of numerical data. Therefore, 128 out of 130 features have numerical values. Two features- ‘custid’ and ‘birthmonth’ are of type object. ‘cardspent’ and ‘card2spent’ are the features providing the credit card spend for the primary and secondary card for a particular customer. The aim is to predict the total card spend, that is, cardspent+card2spent.