Session 12: Tidy Data and Data Types (Solutions Only)

1. Converting Data Types

Q1: Create a `Series` object using the data from the following list, then convert it appropriately to numerical data and compute the sum.

```
l=['Not Available','3.2','5','']
```

```
[9]: l=['Not Available','3.2','5','']
     s=pd.Series(l)
     s
```

```
0    Not Available
1              3.2
2                5
3
dtype: object
```

```
[10]: s=pd.to_numeric(s,errors='coeerce')
      s
```

```
0    NaN
1    3.2
2    5.0
3    NaN
dtype: float64
```

```
[11]: s.sum()
```

```
8.2
```

Q2: Load in the "Marshall_Course_Enrollment_1516_1617.xlsx" file from the classroom schedulling dataset (available on Blackboard and used in session 10), and convert the "Course Suffix" column to numerical format. Then compute the proportion of course suffixes that are 500 or above.

```
[12]: df=pd.read_excel('Marshall_Course_Enrollment_1516_1617.xlsx')
      df['Course Suffix']=pd.to_numeric(df['Course Suffix'],errors='coeerce')
      (df['Course Suffix']>=500).mean()
```

```
0.34080717488789236
```

2. Melting Data

Q3: Run the above code to download the Pew Research Center data on income and religion in the US, and create a DataFrame called "melted" which aggregates the income data into one variable, as shown below.

```
[3]: melted=pew.melt(id_vars='religion',var_name='income',value_name='count')
```

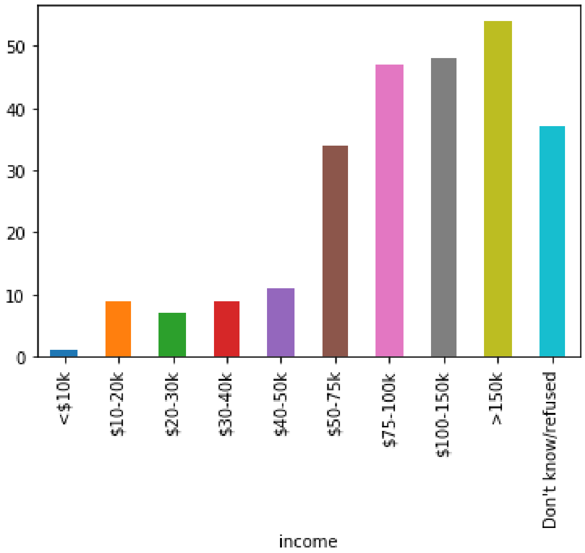
```
[19]: melted.head()
```

```
            religion income  count
0           Agnostic  <$10k     27
1            Atheist  <$10k     12
2           Buddhist  <$10k     27
3           Catholic  <$10k    418
4  Don’t know/refused  <$10k     15
```

Melting the data as above allows you to more easily analyze the income data. For example, the following line plots a histogram of income for Hindus in the US.

```
[5]: melted.query('religion=="Hindu"').plot(x='income',y='count',kind='bar',legend=False)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f0c3f2daf28>
```



3. Pivoting (Un-Melting) Data

Q4: Apply the pivot function on the DataFrame named “melted” you created from Q3, and reset the index so as to get back the original DataFrame.

```
[29]: original=melted.pivot(index='religion',columns='income',values='count').reset_index()
      original.columns.name=''
      original.iloc[:4,:5]
```

```
   religion  $10-20k  $100-150k  $20-30k  $30-40k
0  Agnostic       34        109       60       81
1   Atheist       27         59       37       52
2  Buddhist       21         39       30       34
3  Catholic      617        792      732      670
```