HW1: Datatypes and Wrangling

Hector Corrada Bravo

Feb 2, 2020

Data Types

1) Provide a URL to the dataset.


2) Explain why you chose this dataset.

I am interested in studying how rates of arrests in different parts of Baltimore are related to demographic statistics.

3) What are the entities in this dataset? How many are there?

Entities are specific arrests. There are 104528.

4) How many attributes are there in this dataset?

There are 15 attributes.

5) What is the datatype of each attribute (categorical - ordered or unordered-, numeric - discrete or continuous-, datetime, geolocation, other)? Write a short sentence stating how you determined the type of each attribute. Do this for at least 5 attributes, if your dataset contains more than 10 attributes, choose 10 of them to describe.

<table>
<thead>
<tr>
<th>Num</th>
<th>Name</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>arrest</td>
<td>categorical</td>
<td>Identifier of each arrest, takes values from finite set</td>
</tr>
<tr>
<td>2</td>
<td>age</td>
<td>numeric continuous</td>
<td>Ages are numeric values measured in time units</td>
</tr>
<tr>
<td>3</td>
<td>race</td>
<td>categorical unordered</td>
<td>Can take value from finite set of possible races</td>
</tr>
<tr>
<td>4</td>
<td>sex</td>
<td>categorical unordered</td>
<td>Can take value from finite set of possible sexes</td>
</tr>
<tr>
<td>5</td>
<td>arrestDate</td>
<td>datetime</td>
<td>Specifies date of arrest</td>
</tr>
<tr>
<td>6</td>
<td>arrestTime</td>
<td>datetime</td>
<td>Specifies time of arrest</td>
</tr>
<tr>
<td>7</td>
<td>arrestLocation</td>
<td>other - address</td>
<td>Street address of arrest</td>
</tr>
<tr>
<td>8</td>
<td>incidentOffense</td>
<td>categorical unordered</td>
<td>Can take value from finite set of possible offenses</td>
</tr>
<tr>
<td>9</td>
<td>incidentLocation</td>
<td>other - address</td>
<td>Street address if incident</td>
</tr>
<tr>
<td>10</td>
<td>charge</td>
<td>categorical unordered</td>
<td>Can take value from finite set of possible charges</td>
</tr>
</tbody>
</table>
Write python code that loads the dataset using function `pd.read_csv`. Were you able to load the data successfully? If no, why not?

```python
In [6]:
    import pandas as pd

    url = "http://www.hcbravo.org/IntroDataSci/misc/BPD_Arrests.csv"
    arrest_tab = pd.read_csv(url)
    arrest_tab.head(10)
```

Out[6]:

<table>
<thead>
<tr>
<th>arrest</th>
<th>age</th>
<th>race</th>
<th>sex</th>
<th>arrestDate</th>
<th>arrestTime</th>
<th>arrestLocation</th>
<th>incidentOffense</th>
<th>in</th>
</tr>
</thead>
<tbody>
<tr>
<td>11126858.0</td>
<td>23</td>
<td>B</td>
<td>M</td>
<td>01/01/2011</td>
<td>00:00:00</td>
<td>NaN</td>
<td>Unknown Offense</td>
<td>NaN</td>
</tr>
<tr>
<td>11127013.0</td>
<td>37</td>
<td>B</td>
<td>M</td>
<td>01/01/2011</td>
<td>00:01:00</td>
<td>2000 Wilkens Ave</td>
<td>79-Other</td>
<td>W</td>
</tr>
<tr>
<td>11126887.0</td>
<td>46</td>
<td>B</td>
<td>M</td>
<td>01/01/2011</td>
<td>00:01:00</td>
<td>2800 Mayfield Ave</td>
<td>Unknown Offense</td>
<td>NaN</td>
</tr>
<tr>
<td>11126873.0</td>
<td>50</td>
<td>B</td>
<td>M</td>
<td>01/01/2011</td>
<td>00:04:00</td>
<td>2100 Ashburton St</td>
<td>79-Other</td>
<td>2</td>
</tr>
<tr>
<td>11126968.0</td>
<td>33</td>
<td>B</td>
<td>M</td>
<td>01/01/2011</td>
<td>00:05:00</td>
<td>4000 Wilsby Ave</td>
<td>Unknown Offense</td>
<td>17</td>
</tr>
<tr>
<td>11127041.0</td>
<td>41</td>
<td>B</td>
<td>M</td>
<td>01/01/2011</td>
<td>00:05:00</td>
<td>2900 Spellman Rd</td>
<td>81-Recovered Property</td>
<td>2</td>
</tr>
<tr>
<td>11126932.0</td>
<td>29</td>
<td>B</td>
<td>M</td>
<td>01/01/2011</td>
<td>00:05:00</td>
<td>800 N Monroe St</td>
<td>79-Other</td>
<td>8</td>
</tr>
<tr>
<td>11126940.0</td>
<td>20</td>
<td>W</td>
<td>M</td>
<td>01/01/2011</td>
<td>00:05:00</td>
<td>5200 Moravia Rd</td>
<td>Unknown Offense</td>
<td>NaN</td>
</tr>
<tr>
<td>11127051.0</td>
<td>24</td>
<td>B</td>
<td>M</td>
<td>01/01/2011</td>
<td>00:07:00</td>
<td>2400 Gainsbourgh Ct</td>
<td>54-Armed Person</td>
<td>2</td>
</tr>
<tr>
<td>11127018.0</td>
<td>53</td>
<td>B</td>
<td>M</td>
<td>01/01/2011</td>
<td>00:15:00</td>
<td>3300 Woodland Ave</td>
<td>54-Armed Person</td>
<td>3</td>
</tr>
</tbody>
</table>

Wrangling
1) My pipeline computes average arrest age (ignoring ages <= 0), for each district and writes them in increasing order. It would be useful to see which districts tend to arrest younger individuals.

```python
In [16]: mean_ages = (arrest_tab[['district','age']].
                 query('age > 0').
                 groupby(['district']).
                 agg({'age': 'mean'}).
                 reset_index().
                 sort_values(['age']).

mean_ages
```

Out[16]:

<table>
<thead>
<tr>
<th>district</th>
<th>age</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 NORTHEASTERN</td>
<td>30.431111</td>
</tr>
<tr>
<td>6 SOUTHERN</td>
<td>32.346947</td>
</tr>
<tr>
<td>7 SOUTHWESTERN</td>
<td>32.454487</td>
</tr>
<tr>
<td>5 SOUTHEASTERN</td>
<td>32.515476</td>
</tr>
<tr>
<td>0 CENTRAL</td>
<td>33.056902</td>
</tr>
<tr>
<td>3 NORTHERN</td>
<td>33.128878</td>
</tr>
<tr>
<td>1 EASTERN</td>
<td>34.140232</td>
</tr>
<tr>
<td>8 WESTERN</td>
<td>34.364334</td>
</tr>
<tr>
<td>4 NORTHWESTERN</td>
<td>34.627681</td>
</tr>
</tbody>
</table>

**Plotting**

1) This barplot shows the average arrest age per district (ignoring ages <= 0)
In [17]: from plotnine import *

(ggplot(mean_ages, aes(x='district', y='age')) +
  geom_bar(stat='identity') +
  coord_flip())

  ndistinct = ids.apply(len_unique, axis=0).as_matrix()
/Users/hcorrada/opt/miniconda3/envs/cmsc320/lib/python3.6/site-packages/pandas/core/generic.py:5191: FutureWarning: Attribute 'is_copy' is deprecated and will be removed in a future version.
  return object.__setattr__(self, name, value)
  intervals = data[xminmax].drop_duplicates().as_matrix().flatten()

Out[17]: <ggplot: (291179833)>