HW1: Datatypes and Wrangling

Hector Corrada Bravo

Feb 2, 2020

Data Types

1) Provide a URL to the dataset.

I downloaded my dataset from http://www.hcbravo.org/IntroDataSci/misc/BPD_Arrests.csv (http://www.hcbravo.org/IntroDataSci/misc/BPD_Arrests.csv)

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.

Num	Name	Type	Description
1	`arrest`	categorical	Identifier of each arrest, takes values from finite set
2	`age`	numeric continuous	Ages are numeric values measured in time units
3	`race`	categorical unordered	Can take value from finite set of possible races
4	`sex`	categorical unordered	Can take value from finite set of possible sexes
5	`arrestDate`	datetime	Specifies date of arrest
6	`arrestTime`	datetime	Specifies time of arrest
7	`arrestLocation`	other - address	Street address of arrest
8	`incidentOffense`	categorical unordered	Can take value from finite set of possible offenses
9	`incidentLocation`	other - address	Stree address if incident
10	`charge`	categorical unordered	Can take value from finite set of possible charges

6) Write python code that loads the dataset using function `pd.read_csv`. Were you able to load the data successfully? If no, why not?

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]:

	arrest	age	race	sex	arrestDate	arrestTime	arrestLocation	incidentOffense	in
0	11126858.0	23	B	M	01/01/2011	00:00:00	NaN	Unknown Offense	N
1	11127013.0	37	B	M	01/01/2011	00:01:00	2000 Wilkens Ave	79-Other	W P
2	11126887.0	46	B	M	01/01/2011	00:01:00	2800 Mayfield Ave	Unknown Offense	N
3	11126873.0	50	B	M	01/01/2011	00:04:00	2100 Ashburton St	79-Other	2 S
4	11126968.0	33	B	M	01/01/2011	00:05:00	4000 Wilsby Ave	Unknown Offense	1 S
5	11127041.0	41	B	M	01/01/2011	00:05:00	2900 Spellman Rd	81-Recovered Property	2 R
6	11126932.0	29	B	M	01/01/2011	00:05:00	800 N Monroe St	79-Other	8
7	11126940.0	20	W	M	01/01/2011	00:05:00	5200 Moravia Rd	Unknown Offense	N
8	11127051.0	24	B	M	01/01/2011	00:07:00	2400 Gainsdbourgh Ct	54-Armed Person	2 G
9	11127018.0	53	B	M	01/01/2011	00:15:00	3300 Woodland Ave	54-Armed Person	3 A

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.

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]:

	district	age
2	NORTHEASTERN	30.431111
6	SOUTHERN	32.346947
7	SOUTHWESTERN	32.454487
5	SOUTHEASTERN	32.515476
0	CENTRAL	33.056902
3	NORTHERN	33.128878
1	EASTERN	34.140232
8	WESTERN	34.364334
4	NORTHWESTERN	34.627681

Plotting

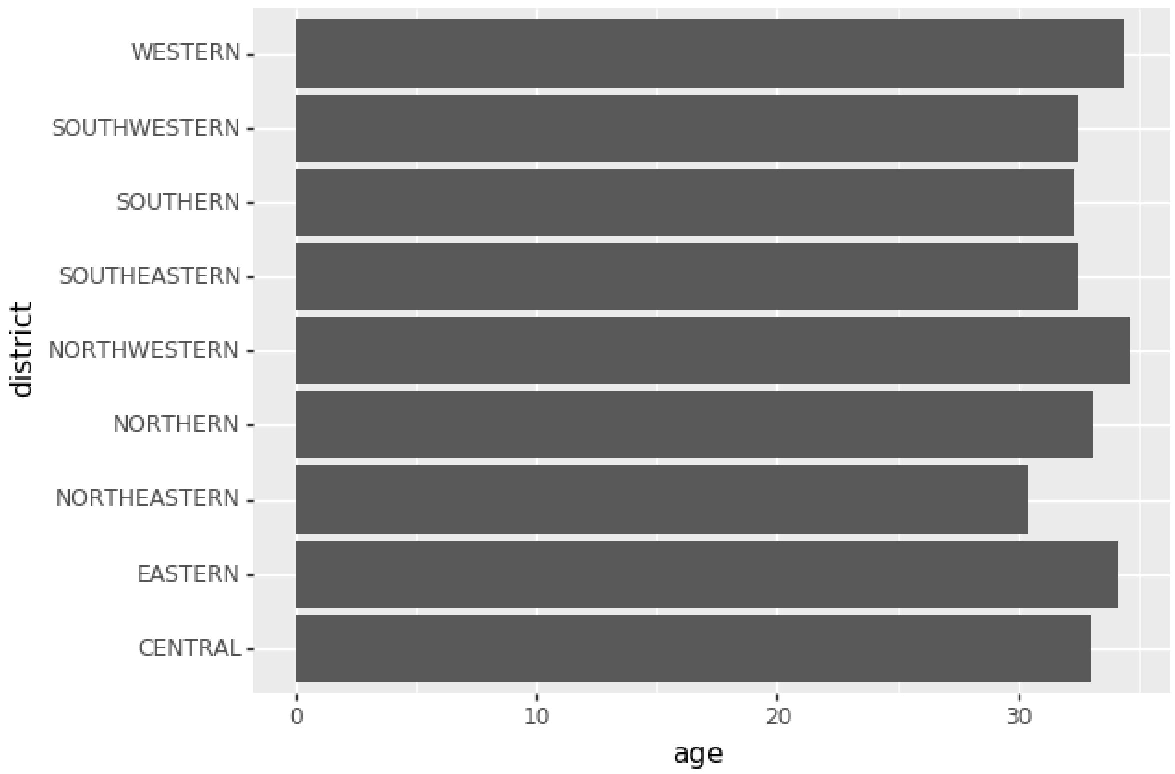
1) This barplot shows the average arrest age per district (ignoring ages <= 0)

In [17]:

```
from plotnine import *

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

```
/Users/hcorrada/opt/miniconda3/envs/cmsc320/lib/python3.6/site-package
s/plotnine/utils.py:284: FutureWarning: Method .as_matrix will be remov
ed in a future version. Use .values instead.
  ndistinct = ids.apply(len_unique, axis=0).as_matrix()
/Users/hcorrada/opt/miniconda3/envs/cmsc320/lib/python3.6/site-package
s/pandas/core/generic.py:5191: FutureWarning: Attribute 'is_copy' is de
precated and will be removed in a future version.
  object.__getattribute__(self, name)
/Users/hcorrada/opt/miniconda3/envs/cmsc320/lib/python3.6/site-package
s/pandas/core/generic.py:5192: FutureWarning: Attribute 'is_copy' is de
precated and will be removed in a future version.
  return object.__setattr__(self, name, value)
/Users/hcorrada/opt/miniconda3/envs/cmsc320/lib/python3.6/site-package
s/plotnine/positions/position.py:188: FutureWarning: Method .as_matrix
will be removed in a future version. Use .values instead.
  intervals = data[xminmax].drop_duplicates().as_matrix().flatten()
```



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

In [ ]: