HW1: Datatypes and Wrangling

Hector Corrada Bravo

2020-01-29

Data types

1) Provide a URL to the dataset.

I downloaded my dataset from 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 R code that loads the dataset using function `read_csv`*. Were you able to load the data successfully? If no, why not?*

```
library(tidyverse)

url <- "http://www.hcbravo.org/IntroDataSci/misc/BPD_Arrests.csv"
arrest_tab <- read_csv(url)
arrest_tab %>% slice(1:10)
```

```
## # A tibble: 10 x 15
##    arrest   age race  sex   arrestDate arrestTime arrestLocation incidentOffense
##     <dbl> <dbl> <chr> <chr> <chr>      <time>     <chr>           <chr>
##  1 1.11e7    23 B     M     01/01/2011 00'00"     <NA>            Unknown Offense
##  2 1.11e7    37 B     M     01/01/2011 01'00"     2000 Wilkens ~ 79-Other
##  3 1.11e7    46 B     M     01/01/2011 01'00"     2800 Mayfield~ Unknown Offense
##  4 1.11e7    50 B     M     01/01/2011 04'00"     2100 Ashburto~ 79-Other
##  5 1.11e7    33 B     M     01/01/2011 05'00"     4000 Wilsby A~ Unknown Offense
##  6 1.11e7    41 B     M     01/01/2011 05'00"     2900 Spellman~ 81-Recovered P~
##  7 1.11e7    29 B     M     01/01/2011 05'00"     800 N Monroe ~ 79-Other
##  8 1.11e7    20 W     M     01/01/2011 05'00"     5200 Moravia ~ Unknown Offense
##  9 1.11e7    24 B     M     01/01/2011 07'00"     2400 Gainsdbo~ 54-Armed Person
## 10 1.11e7    53 B     M     01/01/2011 15'00"     3300 Woodland~ 54-Armed Person
## # ... with 7 more variables: incidentLocation <chr>, charge <chr>,
## #   chargeDescription <chr>, district <chr>, post <dbl>, neighborhood <chr>,
## #   `Location 1` <chr>
```

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.

```
mean_ages <- arrest_tab %>%
  filter(age > 0) %>%
  select(district, age) %>%
  group_by(district) %>%
  summarize(mean_age=mean(age)) %>%
  arrange(mean_age)
mean_ages
```

```
## # A tibble: 10 x 2
##    district     mean_age
##    <chr>           <dbl>
##  1 NORTHEASTERN     30.4
##  2 SOUTHERN         32.3
##  3 SOUTHWESTERN     32.5
##  4 SOUTHEASTERN     32.5
##  5 CENTRAL          33.1
##  6 NORTHERN         33.1
##  7 <NA>             33.4
##  8 EASTERN          34.1
##  9 WESTERN          34.4
## 10 NORTHWESTERN     34.6
```

Plotting

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

```r
mean_ages %>%
  ggplot(aes(x=district, y=mean_age)) +
    geom_bar(stat="identity") +
    coord_flip()
```