Nishant Arora

CMSC 320

HW 1

Data types

1) Provide a URL to the dataset.

Dataset source: https://catalog.data.gov/dataset/most-popular-baby-names-by-sex-and-mothers-ethnic-group-new-york-city-8c742 (https://catalog.data.gov/dataset/most-popular-baby-names-by-sex-and-mothers-ethnic-group-new-york-city-8c742)

2) Explain why you chose this dataset.

I found the data set to be interesting since I visit the city often and was curious. The dataset is clear and concise so I can create a plot easily. The file size also was quite reasonable and not extremely large.

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

The entitites are babies born between 2011 and 2014 in NYC. There are 22,035 entities in the dataset.

4) How many attributes are there in this dataset?

There are 6 attributes: Year of Birth, Gender, Ethnicity, Child's First Name, Count, Rank.

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	`Year of Birth`	numeric-discrete	The year must be 2011, 2012, 2013, or 2014
2	`Gender`	categorical-unordered	Non-numeric, must be from the (unordered) set M, F
3	`Ethnicity`	categorical-unordered	Non-numeric, must be from a finites set of (unordered) ethnicities
4	`Child's First Name`	categorical-unordered	Name is being used to build models
5	`Count`	numeric-discrete	Must be an exact whole number
6	`Rank`	numeric-discrete	Must be an exact whole number

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)

name_tab <- read_csv("/Users/nishant/Desktop/data_science/Most_Popular_Baby_Names_by_Sex_and_Mother_s_Ethnic_Grou
p__New_York_City.csv")
name_tab %>% slice(1:10)
```

```
## # A tibble: 10 x 6
##    `Year of Birth` Gender Ethnicity `Child's First Name` Count  Rank
##              <int> <chr>  <chr>     <chr>                <int> <int>
##  1            2011 FEMALE HISPANIC  GERALDINE               13    75
##  2            2011 FEMALE HISPANIC  GIA                     21    67
##  3            2011 FEMALE HISPANIC  GIANNA                  49    42
##  4            2011 FEMALE HISPANIC  GISELLE                 38    51
##  5            2011 FEMALE HISPANIC  GRACE                   36    53
##  6            2011 FEMALE HISPANIC  GUADALUPE               26    62
##  7            2011 FEMALE HISPANIC  HAILEY                 126     8
##  8            2011 FEMALE HISPANIC  HALEY                   14    74
##  9            2011 FEMALE HISPANIC  HANNAH                  17    71
## 10            2011 FEMALE HISPANIC  HAYLEE                  17    71
```

Wrangling

1. My dataset contains duplicate values, so my pipeline removes duplicate names with the same year, gender and ethnicity, it also displays the top 5 most popular names for Hispanic children born in NYC in 2013.

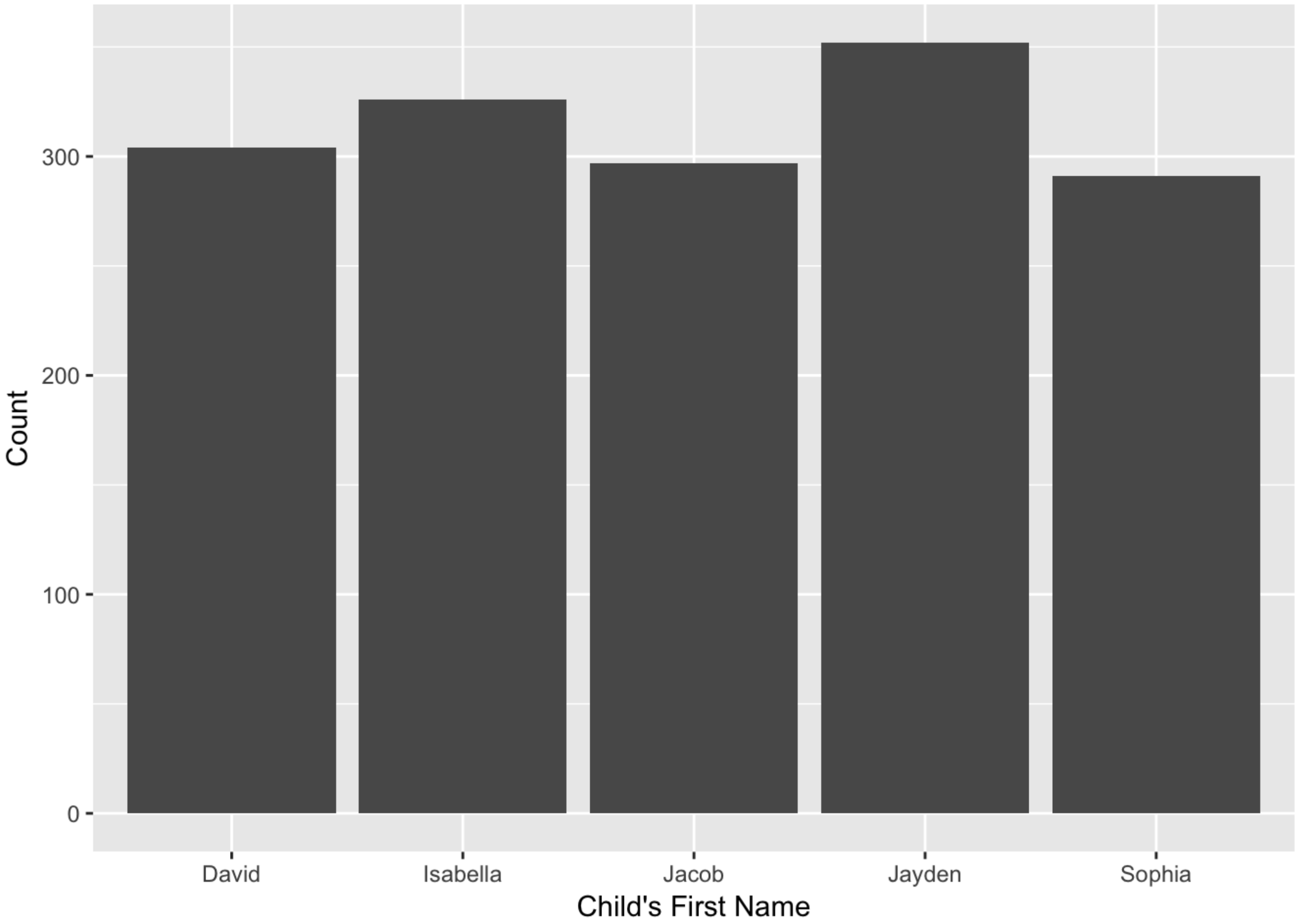
```
top_5_tab <- name_tab %>%
  filter(`Year of Birth` == 2013) %>%
  select(Gender, Ethnicity, `Child's First Name`, Count) %>%
  unique() %>%
  group_by(Ethnicity="HISPANIC") %>%
  arrange(desc(Count)) %>%
  slice(1:5)
top_5_tab
```

```
## # A tibble: 5 x 4
## # Groups:   Ethnicity [1]
##   Gender Ethnicity `Child's First Name` Count
##   <chr>  <chr>     <chr>                <int>
## 1 MALE   HISPANIC  Jayden                 352
## 2 FEMALE HISPANIC  Isabella               326
## 3 MALE   HISPANIC  David                  304
## 4 MALE   HISPANIC  Jacob                  297
## 5 FEMALE HISPANIC  Sophia                 291
```

Plotting

1. This plot shows the 5 most popular Hispanic baby names in NYC in 2013.

```
top_5_tab %>%
  ggplot(aes(x=`Child's First Name`, y=Count)) + geom_bar(stat="identity")
```



```
top_5_tab
```

```
## # A tibble: 5 x 4
## # Groups:   Ethnicity [1]
##   Gender Ethnicity `Child's First Name` Count
##   <chr>  <chr>     <chr>                <int>
## 1 MALE   HISPANIC  Jayden                 352
## 2 FEMALE HISPANIC  Isabella               326
## 3 MALE   HISPANIC  David                  304
## 4 MALE   HISPANIC  Jacob                  297
## 5 FEMALE HISPANIC  Sophia                 291
```