10.1 Final Project Part 2

Brandon Sams

2/13/2020

Data importing and cleaning steps are explained in the text and in the DataCamp exercises.(Tell me why you are doing the data cleaning activities that you perform). Follow a logical process.

```
nyc_dogs <- read.csv("NYC_Dog_Licensing_Dataset.csv")
nyc_dogs$LicenseIssuedDate <- as.Date(nyc_dogs$LicenseIssuedDate)
nyc_dogs$LicenseExpiredDate <- as.Date(nyc_dogs$LicenseExpiredDate)
nyc_dogs <- subset(nyc_dogs,select = -c(Borough))
nyc_dogs$AnimalName[nyc_dogs$AnimalName == "UNKNOWN"] <- NA
nyc_dogs$AnimalName[nyc_dogs$AnimalName == "NAME NOT PROVIDED"] <- NA
nyc_dogs$BreedName[nyc_dogs$BreedName == "Unknown"] <- NA
```

When importing the data originally, the two variables "LicenseIssuedDate" and "LicenseExpiredDate" were imported as factors, instead of dates. I changed these to dates, because that is what they are. I also used excel to change the date format to be of the form yyyy-mm-dd prior to importing the data into R.

I also made the decision to drop the Borough column entirely, as there were no values for that variable in that data set. Every entry was blank. I could possibly deduce what the borough is by the zip code later, if that becomes a piece of relevant information.

I also noticed that a good amount of the AnimalName was "UNKNOWN" or "NAME NOT PROVIDED". I think it reasonable to say that these are missing values, rather than actual dog names. I set these values to NA. I also did a similar treatment with the BreedName as well.

The purpose of the "Extract.Year" variable is unclear, but I left it in there in the event that it becomes useful later.

With a clean dataset, show what the final data set looks like. However, do not print off a data frame with 200+ rows; show me the data in the most condensed form possible.

```
head(nyc_dogs,5)
```

```
##   RowNumber AnimalName AnimalGender AnimalBirthMonth
## 1         1      PAIGE            F             2014
## 2         2       YOGI            M             2010
## 3         3        ALI            M             2014
## 4         4      QUEEN            F             2013
## 5         5       LOLA            F             2009
##                               BreedName ZipCode LicenseIssuedDate
## 1 American Pit Bull Mix / Pit Bull Mix   10035        2014-09-12
## 2                                Boxer   10465        2014-09-12
## 3                              Basenji   10013        2014-09-12
## 4                    Akita Crossbreed   10013        2014-09-12
## 5                              Maltese   10028        2014-09-12
##   LicenseExpiredDate Extract.Year
## 1         2017-09-12         2016
```

```
## 2          2017-10-02          2016
## 3          2019-09-12          2016
## 4          2017-09-12          2016
## 5          2017-10-09          2016
```

What do you not know how to do right now that you need to learn to import and cleanup your dataset?

I know some dog breeds, but certainly not all of them. So when I am doing anaylsis, I would like the dog breeds which are actually the same to have the same value. This is not necessarily the case. I do not know of a way to do a spell check, for example. This would correct incorrectly entered data, if there is any.

I also do not know of a way to check other data forms for validity. There may be erroneous data throughout the entire dataset, and I would not have a way of knowing.

Discuss how you plan to uncover new information in the data that is not self-evident. I am importing the data into MySQL to slice and dice the data a little more easily. Yes, I am pretty sure you can do that in R, but MySQL has a lot of features that help to make that just a bit easier. The syntax for finding distinct values is helpful to see if there are any values that just do not make any sense.

What are different ways you could look at this data to answer the questions you want to answer? I think it would be interesting to see if this data set as a collection of registrations. A dog could, from my understanding, appear in the data set multiple times, as the owner continues to register them after the previous registration expires. Perhaps there is a way that a new data frame could be created that attempts to represent an individual pet, and the number of times that it has been registred.

Do you plan to slice and dice the data in different ways, create new variables, or join separate data frames to create new summary information? Explain. The ZipCode information is super interesting, but it is hard to parse in its current state. A zip code does not tell me much in itself, but I think it would be good to find a way to map that a more specific physical location.

I found this website that will tell you what Borough a Zip Code belongs to, but also which Neighborhood it belongs to, in that Borough.

Another way to look at the data is by looking at the difference between when their registration starts and when it ends. It is not a consistent amount, so a pattern may arise that underrepresents dogs that have long registration periods.

How could you summarize your data to answer key questions? I suppose the best way to summarize this data is with several graphs and charts. Perhaps a histogram of dog breeds. Most frequent dog names by year and gender.

I could also attempt to break down dog registrations per capita by year.

I would love to see a heatmap for NYC that shows regions that tend to have more dog owners. I know that one of the joys of owning a dog is meeting other dog owners.

What types of plots and tables will help you to illustrate the findings to your questions? Ensure that all graph plots have axis titles, legend if necessary, scales are appropriate, appropriate geoms used, etc.). I think that the most helpful graphs will be ones that show how frequently a certain breed is present. I think it would also be helpful to show which zip codes have the most dogs registered in them. These would, of course, need helpful titles and legends and such. I would probably use histograms.

What do you not know how to do right now that you need to learn to answer your questions? I do not know how to make a map of NYC that shows how many dogs are (or have been?) registered in that area. There may be a package that makes that easier.

Do you plan on incorporating any machine learning techniques to answer your research questions? Explain. I do not plan on incorporating any machine learning at this point. It seems that is good for making predictions, but this data set does not seem to necessitate that.