DSC 520 Final Project

Name: Jonathan Lawrence

Date: 6/1/2019

Section 1

- Explain what your interests are in the data sets identified.

  My interest is to gain insights about animal-related incidents in the DFW area. I'd like to analyze the type of animals that they take in the most, see if there is any correlation with the time of the year, and possibly a few other things.

- What is the target audience for this research?

  My mother-in-law is the target audience for my research. She volunteers and financially supports an animal rescue group here in the Dallas area. I believe the information I give her could help to identify areas of interest for her among other things.

- Identify the Packages that are needed for your project.

  ggplot: visualizations
  lubridate: time/date
  pastecs: descriptive statistics
  RMarkdown: Creating a PDF document

- Original source where the data was obtained is cited and, if possible, hyperlinked.

  Dataset: Dallas Animal Shelter Data
  https://www.opendatanetwork.com/dataset/www.dallasopendata.com/7h2m-3um5

- Source data is thoroughly explained (i.e. what was the original purpose of the data, when was it collected, how many variables did the original have, explain any peculiarities of the source data such as how missing values are recorded, or how data was imputed, etc.).

  The original purpose of this data was to help citizens better understanding the operational processes that the shelter personnel perform daily for the animals and citizens of the City of Dallas. The dataset is updated daily starting from October 1st 2017 to present. There are 34 variables. Missing values are left blank, so no imputation occurred.

Section 2

- Provide an introduction that explains the problem statement you are addressing. Why would someone be interested in this?
    - **"Where does the Dallas Animal Shelter have the greatest need?" Many people donate to animal shelters, but a sparse few truly understand how to make the biggest impact. By investigating the shelter's data for all of 2018, people can draw insights about the type of incidents that the shelter deals with in order to learn where best to apply their support and/or resources.**
- Provide a concise explanation of how you plan to address this problem statement
    - **I plan to address it by learning as much as possible about the animal incidents at one of Dallas' largest animal shelters.**
- Discuss how your proposed approach will address (fully or partially) this problem.
    - **My approach will address this problem by identifying some key information to help determine what kinds of animals the shelter deals most, the animal's condition, and various other factors that could reveal where the shelter's biggest needs are.**
- List at least 6 research questions you aim to answer.
    - **What type of animal receives the most incidents at the shelter?**
    - **What breed of dog has the most incidents?**
    - **What type of intake does the shelter see the most?**
    - **What type of outtake does the shelter see the most?**
    - **During which time of the year does the shelter receive the most animals?**
    - **Are the majority of the animals taken in healthy or sick?**
- Explain how your analysis may help the consumer of your research findings (recall you target audience from Section 1).
    - **Since my mother-in-law helps to support an animal shelter, she may be able to use this information in order to apply her support where it is most needed. For example, if I discover that most of the animals taken in are sick, she may allocate her resources to providing more medication. Learning about the type of animals taken in could help her adjust the types of items donated to an amount proportionate to the animal intake.**
- What types of plots and tables will help you to illustrate the findings to your research questions?
    - **I think histograms will be helpful to show the time most animal incidents occur. Meanwhile, I think bar graphs will be a main focus of my data since most of the research questions deal with determining the majority factor from different parts of the data.**
- What do you not know how to do right now that you need to learn to answer your research questions?
    - **Right now I need to determine the best way to visually represent my questions including coloring. I have ideas of how to do that, but I am not aware of the best way to visualize it.**

Section 3

- 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) and follow a logical process.

  For the first step in my cleaning process, I'll check which variables include NA/blank values and ensure that all blank values are replaced with NA. I'll then removed the NA values using listwise deletion. I'll also be using the filter feature in Excel to detect any abnormal responses or typos such as "D" instead of "DOG". Where able, I'll clump answers like that together in my analysis.

- 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.

Cleaned dataset (condensed)

```
##     Animal.Id        Animal.Type          Animal.Breed
##  A0979270:    8   BIRD     :  353   DOMESTIC SH  : 8153
##  A1022242:    8   CAT      : 9176   PIT BULL     : 6416
##  A1027417:    8   D        :    0   CHIHUAHUA SH : 3866
##  A1022243:    7   DOG      :28056   LABRADOR RETR: 3581
##  A1023489:    7   LIVESTOCK:   23   GERM SHEPHERD: 3525
##  A1039510:    7   WILDLIFE : 1092   CAIRN TERRIER:  884
##  (Other) :38655                     (Other)      :12275
##    Kennel.Number       Kennel.Status   Tag.Type        Activity.Number
##  RECEIVING: 2904   UNAVAILABLE:16818   Mode:logical   A18-129535:   68
##  FOSTER   : 2036   IMPOUNDED  :11632   NA's:38700     A18-134873:   57
##  RTO FIELD: 2028   LAB        : 6222                  A18-105700:   54
##  LAB 01   : 1261   AVAILABLE  : 3299                  A18-111539:   44
##  STAR 5   :  669   WILDLIFE   :  336                  A18-106373:   43
##  PSDOG 01 :  526   PRE-LAB    :  257                  (Other)   :21605
##  (Other)  :29276   (Other)    :  136                  NA's      :16829
##  Activity.Sequence    Source.Id      Census.Tract    Council.District
##  Min.   : 0.000    P0000000:11388   20500  : 3040   6      : 6398
##  1st Qu.: 1.000    P9991763:  223   17004  :  695   4      : 5263
##  Median : 1.000    P9999999:  217   8400   :  685   8      : 5173
##  Mean   : 1.027    P9991755:  206   17102  :  609   5      : 4275
##  3rd Qu.: 1.000    P0851546:  127   11602  :  601   7      : 3727
##  Max.   :30.000    P0731334:   98   (Other):33068   (Other):13862
##                    (Other) :26441   NA's   :    2   NA's   :    2
##            Intake.Type           Intake.Subtype   Intake.Total
##  CONFISCATED    : 1462   AT LARGE      :18434   Min.   :1
##  FOSTER         : 1839   GENERAL       :10956   1st Qu.:1
##  OWNER SURRENDER:10548   CONFINED      : 3774   Median :1
##  STRAY          :23636   POSSIBLY OWNED: 1224   Mean   :1
##  TRANSFER       :  108   QUARANTINE    :  956   3rd Qu.:1
##  TREATMENT      :  162   RETURN30      :  856   Max.   :1
##  WILDLIFE       :  945   (Other)       : 2500
##            Reason         Staff.Id        Intake.Date       Intake.Time
##  TOO MANY     :  939   SC1704 :  884   5/19/2018:  201   12:00:00:  126
##  OWNER PROBLEM:  817   LL     :  795   6/23/2018:  191   12:12:00:  122
##  MOVE         :  791   KV1734 :  743   7/17/2018:  188   11:05:00:  117
##  NO TIME      :  545   DB1715 :  734   6/6/2018 :  178   11:07:00:  114
##  LANDLORD     :  515   CS1750 :  712   7/11/2018:  176   12:22:00:  111
##  (Other)      : 4017   YL1695 :  639   9/13/2018:  176   11:04:00:  106
##  NA's         :31076   (Other):34193   (Other)  :37590   (Other) :38004
##       Due.Out                                    Intake.Condition
##  6/13/2018:  222   TREATABLE REHABILITABLE NON-CONTAGIOUS:30729
##  5/19/2018:  206   UNHEALTHY UNTREATABLE NON-CONTAGIOUS  : 3209
##  7/17/2018:  186   TREATABLE MANAGEABLE NON-CONTAGIOUS   : 3016
##  6/22/2018:  178   HEALTHY                               :  604
##  6/30/2018:  173   UNHEALTHY UNTREATABLE CONTAGIOUS      :  604
##  6/27/2018:  168   TREATABLE REHABILITABLE CONTAGIOUS    :  346
##  (Other)  :37567   (Other)                               :  192
##                    Hold.Request             Outcome.Type
##  ADOP RESCU              : 8241   ADOPTION         :13480
##  RESCU ONLY              : 4176   TRANSFER         : 7617
##  ADOPTION                : 4086   RETURNED TO OWNER: 7506
##  EVERYDAY ADOPTION CENTER: 2232   EUTHANIZED       : 6790
##  MEDICAL                 : 1472   FOSTER           : 1895
##  (Other)                 : 5564   WILDLIFE         :  473
##  NA's                    :12929   (Other)          :  939
##   Outcome.Subtype     Outcome.Date     Outcome.Time      Receipt.Number
##  WALK IN  :13872   8/18/2018:  298   0:00:00 :  864   R18-533381:    8
##  OTHER    : 5957   1/24/2018:  188   18:00:00:  124   R18-530615:    3
##  FIELD    : 4023   9/13/2018:  188   17:19:00:  115   R18-530979:    3
##  PROMOTION: 2572   7/9/2018 :  183   17:17:00:  114   R18-531079:    3
##  HUMANE   : 1854   6/20/2018:  181   17:52:00:  114   R18-523527:    2
##  BEHAVIOR : 1759   (Other)  :37660   17:02:00:  113   (Other)   :16310
##  (Other)  : 8663   NA's     :    2   (Other) :37256   NA's      :22371
##     Impound.Number  Service.Request.Number
##  K15-309916:    1   1800252670:    9
##  K17-395525:    1   B         :    7
##  K17-401589:    1   A         :    4
##  K17-403869:    1   1800061265:    3
##  K17-403964:    1   1800717174:    3
##  K18-______:    1   (Other)   :  203
##  (Other)   :38694   NA's      :38471
##                               Outcome.Condition
##  TREATABLE REHABILITABLE NON-CONTAGIOUS:27788
##  UNHEALTHY UNTREATABLE NON-CONTAGIOUS  : 5006
##  TREATABLE MANAGEABLE NON-CONTAGIOUS   : 2873
##  UNHEALTHY UNTREATABLE CONTAGIOUS      : 1289
##  HEALTHY                               :  612
##  (Other)                               : 1036
##  NA's                                  :   96
##                     Chip.Status             Animal.Origin
##  SCAN CHIP                : 8919   FIELD           :13548
##  SCAN NO CHIP             :27345   OVER THE COUNTER:17528
##  UNABLE TO SCAN           : 2435   SWEEP           : 7623
##  WILDLIFE - UNABEL TO SCAN:    0   NA's            :    1
##  NA's                     :    1
```

```
## 
## 
##        Additional.Information      Month          Year
##  ADOPTED           : 4017      JUN.2018: 3766   FY2018:28973
##  TAGGED            : 3033      JUL.2018: 3765   FY2019: 9727
##  ADOPT PENDING     : 1082      DEC.2018: 3523
##  RETURNED TO OWNER:  886      AUG.2018: 3515
##  FOSTER            :  732      MAY.2018: 3411
##  (Other)           :13435      SEP.2018: 3341
##  NA's              :15515      (Other) :17379
```

```
head(dat)
```

```
##   Animal.Id Animal.Type  Animal.Breed Kennel.Number Kennel.Status Tag.Type
## 1  A0767064         DOG         BOXER       LFD 080     IMPOUNDED       NA
## 2  A1030017         CAT   DOMESTIC SH        FOSTER     IMPOUNDED       NA
## 3  A1024088         DOG   POODLE STND     RTO FIELD     IMPOUNDED       NA
## 4  A1014535         DOG  CHIHUAHUA SH      PSDOG 01     AVAILABLE       NA
## 5  A1012414         DOG GERM SHEPHERD     RTO FIELD     IMPOUNDED       NA
## 6  A1018447         DOG COLLIE SMOOTH        STAR 4     AVAILABLE       NA
##   Activity.Number Activity.Sequence Source.Id Census.Tract
## 1            <NA>                 1  P0815009         5300
## 2            <NA>                 1  P0000000        11701
## 3      A18-100531                 1  P0000000         9607
## 4      A17-084598                 1  P0812059         1204
## 5      A17-081201                 1  P0000000         9303
## 6      A18-090640                 1  P0817785        12702
##   Council.District     Intake.Type Intake.Subtype Intake.Total Reason
## 1                1           STRAY POSSIBLY OWNED            1   <NA>
## 2                5           STRAY       AT LARGE            1   <NA>
## 3               13           STRAY       AT LARGE            1   <NA>
## 4                2 OWNER SURRENDER        GENERAL            1   <NA>
## 5                5           STRAY       AT LARGE            1   <NA>
## 6                9           STRAY POSSIBLY OWNED            1   <NA>
##   Staff.Id Intake.Date Intake.Time    Due.Out
## 1    BW/LW  12/11/2017    17:34:00 12/22/2017
## 2   AR1577   5/16/2018    12:10:00  5/20/2018
## 3 JAS 1719   3/15/2018    11:11:00  3/15/2018
## 4       MB  11/16/2017    10:45:00 11/16/2017
## 5      SEC  10/25/2017     8:22:00 10/29/2017
## 6   LH1714    1/6/2018    11:59:00  1/17/2018
##                          Intake.Condition             Hold.Request
## 1 TREATABLE REHABILITABLE NON-CONTAGIOUS              HOLD NOTIFY
## 2 TREATABLE REHABILITABLE NON-CONTAGIOUS                     <NA>
## 3 TREATABLE REHABILITABLE NON-CONTAGIOUS                     <NA>
## 4    TREATABLE MANAGEABLE NON-CONTAGIOUS EVERYDAY ADOPTION CENTER
## 5 TREATABLE REHABILITABLE NON-CONTAGIOUS                     <NA>
## 6 TREATABLE REHABILITABLE NON-CONTAGIOUS               RESCU ADOP
##        Outcome.Type Outcome.Subtype Outcome.Date Outcome.Time
## 1 RETURNED TO OWNER         WALK IN   12/12/2017     11:54:00
## 2            FOSTER           STAFF    5/16/2018     23:40:00
## 3 RETURNED TO OWNER           FIELD    3/15/2018     11:41:00
## 4          ADOPTION         WALK IN   11/19/2017     10:39:00
## 5 RETURNED TO OWNER           FIELD   10/25/2017      8:28:00
## 6          ADOPTION       PROMOTION    1/24/2018     15:56:00
##   Receipt.Number Impound.Number Service.Request.Number
## 1     R17-520505     K17-402459                   <NA>
## 2           <NA>     K18-417500                   <NA>
## 3           <NA>     K18-411002                   <NA>
## 4     R17-519399     K17-400236                   <NA>
## 5           <NA>     K17-398008                   <NA>
## 6     R18-522816     K18-404643                   <NA>
##                        Outcome.Condition  Chip.Status    Animal.Origin
## 1 TREATABLE REHABILITABLE NON-CONTAGIOUS    SCAN CHIP OVER THE COUNTER
## 2   UNHEALTHY UNTREATABLE NON-CONTAGIOUS SCAN NO CHIP            SWEEP
## 3 TREATABLE REHABILITABLE NON-CONTAGIOUS SCAN NO CHIP            SWEEP
## 4    TREATABLE MANAGEABLE NON-CONTAGIOUS SCAN NO CHIP            FIELD
## 5 TREATABLE REHABILITABLE NON-CONTAGIOUS SCAN NO CHIP            SWEEP
## 6 TREATABLE REHABILITABLE NON-CONTAGIOUS    SCAN CHIP            FIELD
##   Additional.Information    Month   Year
## 1      RETURNED TO OWNER DEC.2017 FY2018
## 2                   <NA> MAY.2018 FY2018
## 3                   <NA> MAR.2018 FY2018
## 4             SNN DALLAS NOV.2017 FY2018
## 5                   <NA> OCT.2017 FY2018
## 6          FREE ADOPTION JAN.2018 FY2018
```

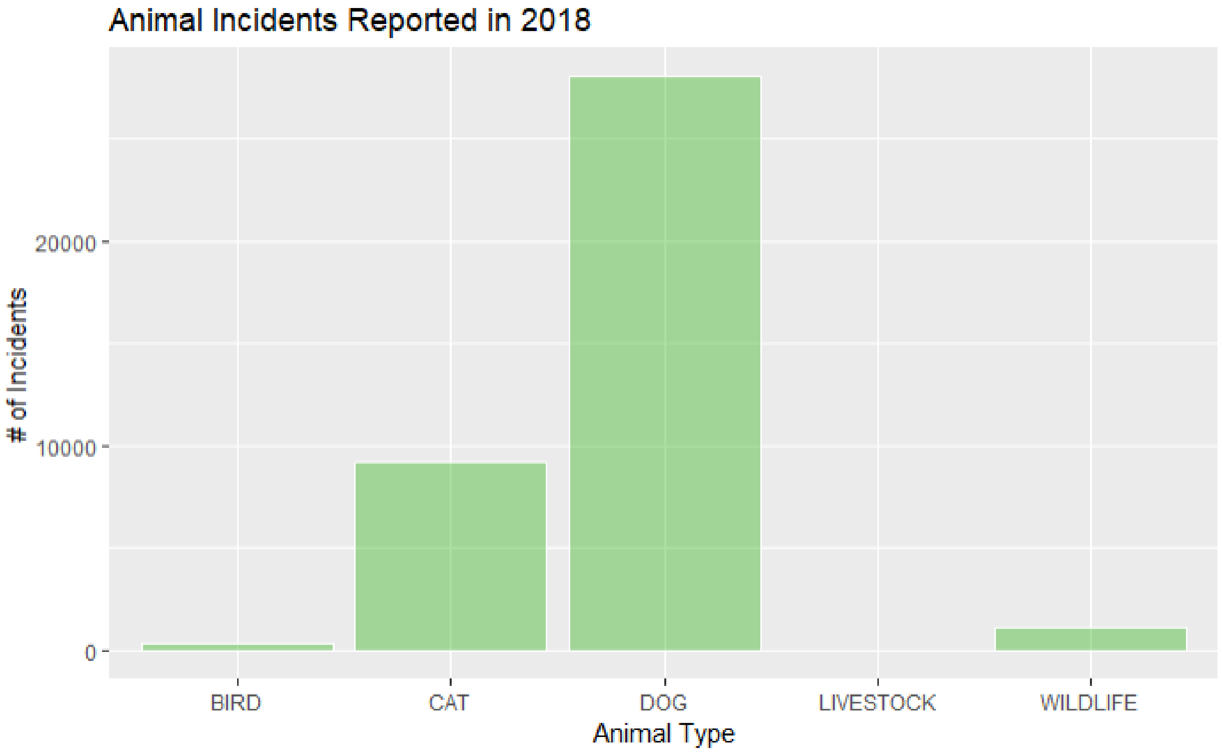
- What do you not know how to do right now that you need to learn to import and cleanup your dataset?
    - **Memorizing useful R packages that I will need for my plots**
    - **Inherently knowing how to transform my data (e.g. combine columns, work with dates)**
    - **Removing outliers and bias**

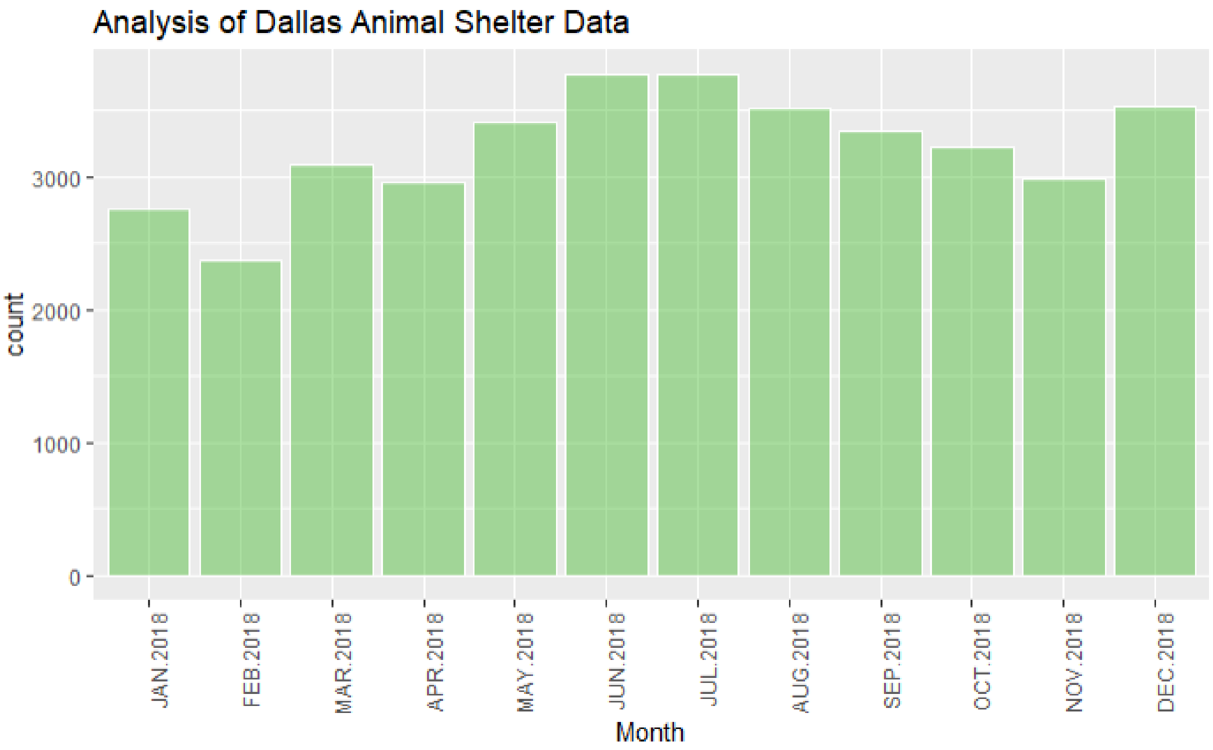
Section 4

- Discuss how you plan to uncover new information in the data that is not self-evident.
    - **I plan to discover which months the animal shelter takes in the most dogs.**
- What are different ways you could look at this data to answer the questions you want to answer?
    - **I can plot variables over time to determine problem areas for the Dallas animal shelter. For example, I can determine if the number of dogs given up over-the-counter is increasing faster than dogs retrieved from the field.**
- 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.
    - **A lot of my data will be subset in various ways. For example, I plan to locate any problematic locations in the data where typos or misspellings exist and combine them**

to avoid issues. In some cases, I might also combine all animals into one result so that we can determine overall quantities regardless of species.

- How could you summarize your data to answer key questions?
    - **The data I've acquired can be summarized by plotting it against time. Determining various statistics at certain points in time will help to find trends in the data from 2018 and might even help predict what we can expect for the rest of 2019.**
- 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'd like to use some standard plots over time which will help me to illustrate trends based on species as well as insights about particular months when most animals are turned in. Example:**





- What do you not know how to do right now that you need to learn to answer your questions?
  - **I need to learn how to write some of the code to do what I want. This will come with study and time. I know all of my questions, but not necessarily how to work them into code. I also need more time practicing general guidelines for writing ggplot code. Finally, I need to practice creating graphs that are not the typical bar graph or histogram, and also colorizing them.**
- Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.
  - **At this time, my practical knowledge of machine learning techniques is very slim. There might be some use cases down the road, but right now I am not planning to implement machine learning.**

Suggestion from the course professor: Some additional questions you may want to consider asking yourself as you work through this section of the project:

1. What features could you filter on? **I can filter on species that don't help me answer my questions such as turtles which I have spotted in the data.**
2. How could arranging your data in different ways help? **I am not aware of any reason to rearrange the data considering how it is already in a nice arrangement and can do everything I need it to.**
3. Can you reduce your data by selecting only certain variables? **Yes I plan to ignore a lot of the variables because they are either open-end responses that are difficult to draw insights from, or they are very obscure facts that do not contribute to answering my questions.**
4. Could creating new variables add new insights? **I'm sure it could, but at this time I believe the data provides all of the variables I will need.**

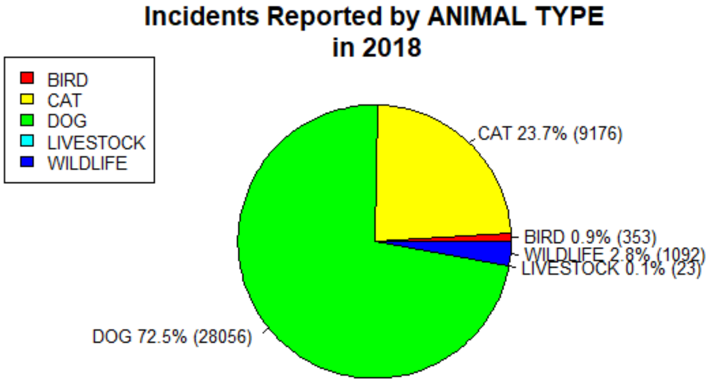
5. Could summary statistics at different categorical levels tell you more? **It's possible, but I will need more time to determine if it is necessary to answer my questions.**
6. How can you incorporate the pipe (%>%) operator to make your code more efficient? **I used the pipe operator to filter out dog breeds with less than 500 reported incidents in the data.**

Section 5 Summary

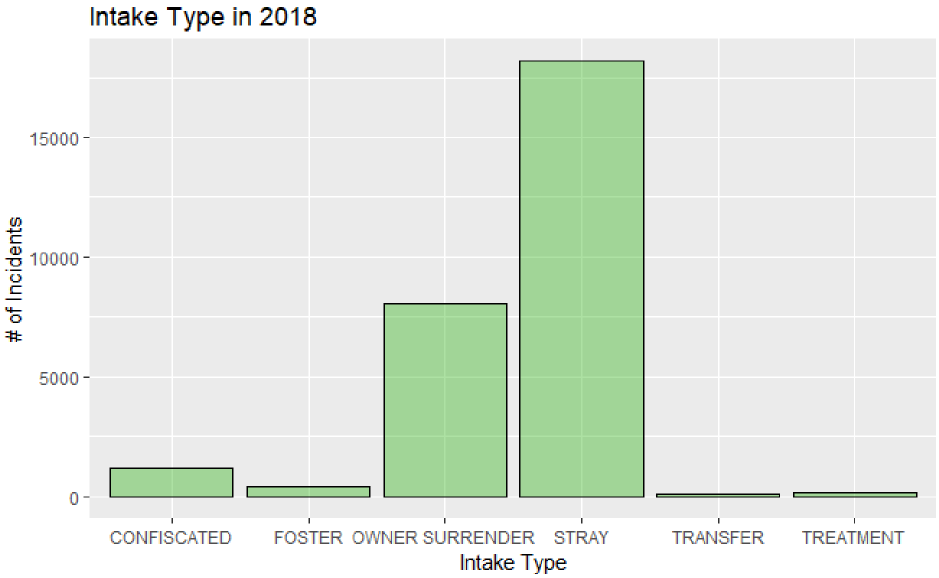
- Overall, write a coherent narrative that tells a story with the data as you complete this section.
- Summarize the problem statement you addressed.
- Summarize how you addressed this problem statement (the data used and the methodology employed).
- Summarize the interesting insights that your analysis provided.
- Summarize the implications to the consumer (target audience) of your analysis.
- Discuss the limitations of your analysis and how you, or someone else, could improve or build on it.
- In addition, submit your completed Project using R Markdown or provide a link to where it can also be downloaded from and/or viewed.

"Where does the Dallas Animal Shelter have the greatest need?" Animal shelters receive aid through donations made by generous people who want to assist with the care for ownerless animals. People can donate food, medical supplies, and their time through volunteer activities in order to promote healthy lives for the animals at the shelter. But for the strong of heart, is there a way to ascertain where the shelter's most dire need is in order to have the greatest impact? My mother-in-law, Molly, is a big supporter of the Dallas Animal Shelter and wants to go above and beyond the normal donor. By investigating 2018 data of approximately 40,000 animals received at the Dallas Animal Shelter, we can draw insights in order to learn where it would be best for her to apply her support and/or resources.

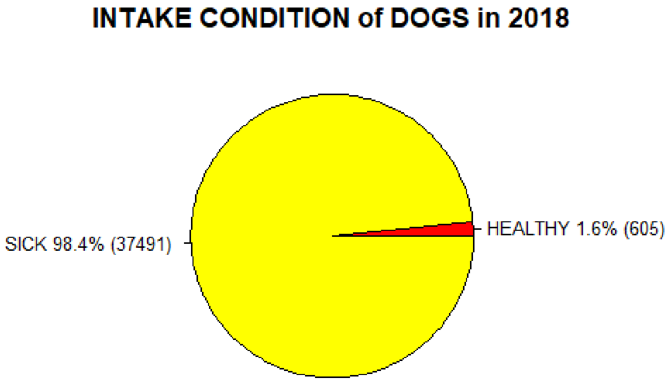
To begin, I isolated the area of greatest need by asking a few questions about the animals going through the shelter. This included the type of animal most seen, its condition, time of year, and various other factors. The data revealed that dogs made up a vast majority of the animals admitted to the shelter for 2018, more than twice the number of cats which was the second highest.



From this first analysis we can assume that the shelter has a need for donations geared towards helping dogs. But we can go further by asking the question "How can we help the dogs?" We can answer this by determining where they came from and what their condition was when they entered the shelter. The data shows that strays were more common in 2018 than all other types of intake combined.

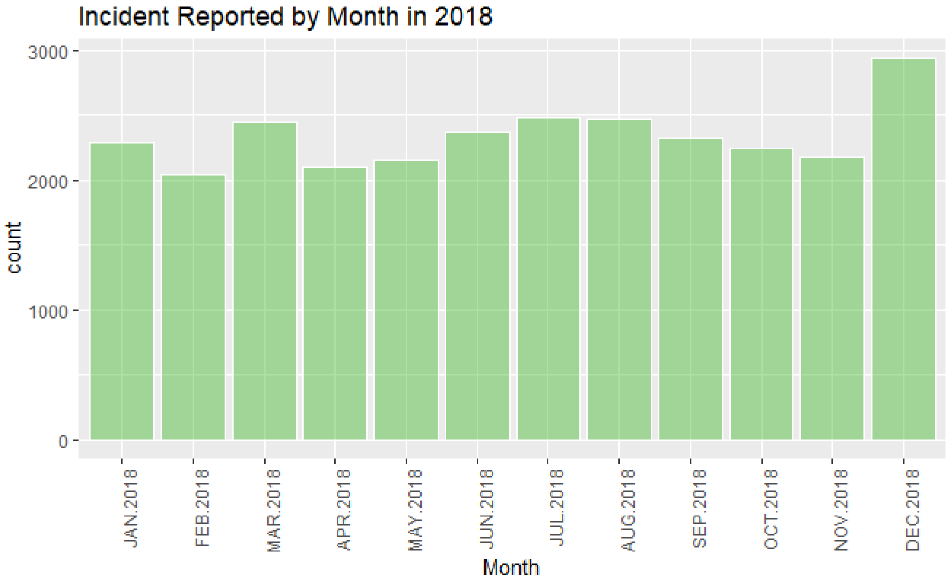


Given that most dogs at the shelter were strays, this begged the question of whether or not the dogs were received healthy or sick. My analysis revealed that a shocking 98.4% of all dogs admitted were sick with some type of medical illness.



The quantity of dogs and their likelihood of being sick upon receipt, tells us that the shelter is likely to have a dire need for donations geared towards sick dogs. However, this analysis is limited by a few factors. For example, just because dogs are the most common animal received doesn't mean that dog food is their greatest need. It could be that they have a surplus of dog food, and a lack of cat food. Having the chance to look at the donation data and the shelter's inventory could help to fine-tune this analysis. Unfortunately, the Dallas Animal Shelter does not provide the public with access to this data.

Now that we know <u>how</u> Molly can help, we want to know <u>when</u> the best time for Molly to donate would be. By examining the number of dogs per month I was able to determine that the month of December had the highest intake of dogs in 2018. The data also showed that the number of intakes rose slightly during summer months. The latter could be due to the higher likelihood of pet owners moving during the summer and leaving their pets behind. This analysis could be improved by comparing the rate of intakes per month by the rate of residents leaving the city, but we do not have access to that data for the Dallas area in order to make a proper correlation.



In summary, the greatest area of need for the Dallas Animal Shelter in 2018 was donations for sick dogs during the month of December. Molly can use this information to tailor her 2019 donations.