Apoorva Balasubramanian: 1550424
Rishabh Malhotra: 1526907
Ashish Mahajan: 1526229

PROJECT PROPOSAL

1. OVERVIEW:

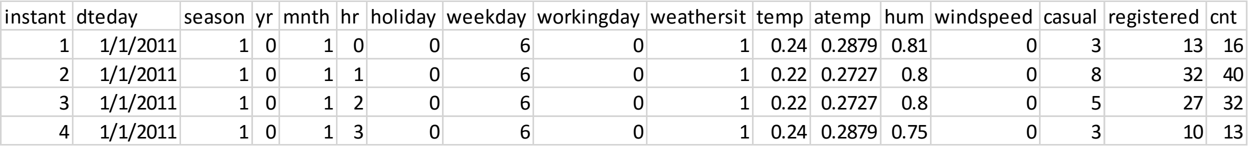
The goal of this project is to analyze data pattern in the given dataset and predict outcomes for new data by using various machine learning techniques.

2. DATASETS & ANALYSIS:

2.1. Dataset 1→ Bike Sharing Demand Data:

Bike sharing systems are new generation of traditional bike rentals where whole process from membership, rental and return back has become automatic. Through these systems, user is able to easily rent a bike from a particular position and return back at another position. This dataset contains the hourly and daily count of rental bikes between years 2011 and 2012 in Capital bike share system with the corresponding weather and seasonal information.

Following is the snapshot of the dataset:



instant	dteday	season	yr	mnth	hr	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
1	1/1/2011	1	0	1	0	0	6	0	1	0.24	0.2879	0.81	0	3	13	16
2	1/1/2011	1	0	1	1	0	6	0	1	0.22	0.2727	0.8	0	8	32	40
3	1/1/2011	1	0	1	2	0	6	0	1	0.22	0.2727	0.8	0	5	27	32
4	1/1/2011	1	0	1	3	0	6	0	1	0.24	0.2879	0.75	0	3	10	13

2.1.1. SUMMARY:

- *Number of Instances* = 17389
- *Number of Attributes* = 16

2.1.2. ATTRIBUTES:

- *Instant* → Record Index
- *Dteday* → Date
- *Season* → season

1: Springer	2: Summer	3: Fall	4: Winter

- *Yr* →year

0: 2011	1: 2012

- *mnth* : month

1: January to 12: December

- *Hr*: hour

0 to 23

- *Holiday*: weather day is holiday or not.

- *Weekday*: day of the week

Working Day: 1 If day is neither weekend nor holiday	Otherwise : 0

- *Weathersit* :

1	Clear, Few clouds, Partly cloudy, Partly cloudy
2	Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
3	Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
4	Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

- *Temp*:
  Normalized temperature in Celsius. The values are derived via→

$$\frac{(t - t\_\mathrm{min})}{(t\_\,\mathrm{max} - t\_\mathrm{min})}$$

  Where t_min = -8 and t_max = +39 (only in hourly scale)

- *Atemp*: Normalized feeling temperature in Celsius. The values are derived via→

$$\frac{(t - t\_\mathrm{min})}{(t\_\,\mathrm{max} - t\_\mathrm{min})}$$

  Where t_min=-16, t_max=+50 (only in hourly scale)

- *Hum*:
  Normalized humidity. The values are divided to 100 (max)

- *Windspeed*:
  Normalized wind speed. The values are divided to 67 (max)

- *Casual*: count of casual users.
- *Registered*: count of registered users.
- *Cnt*: count of total rental bikes including both casual and registered.

2.1.3. MACHINE LEARNING QUESTIONS:

Question 1→
For a given time, day, weather situation and climatic conditions how many casual users will be utilizing the bike sharing service?

Question 2→
How many registered users will be utilizing the bike sharing service for a given time, day, weather situation and climatic conditions?

Question 3→
In the next hour how many users are expected to use the bikes from bike sharing system?

Question 4→
Forecast the number of bikes that will be required to meet the demand the upcoming season so that we can plan the inventory accordingly?

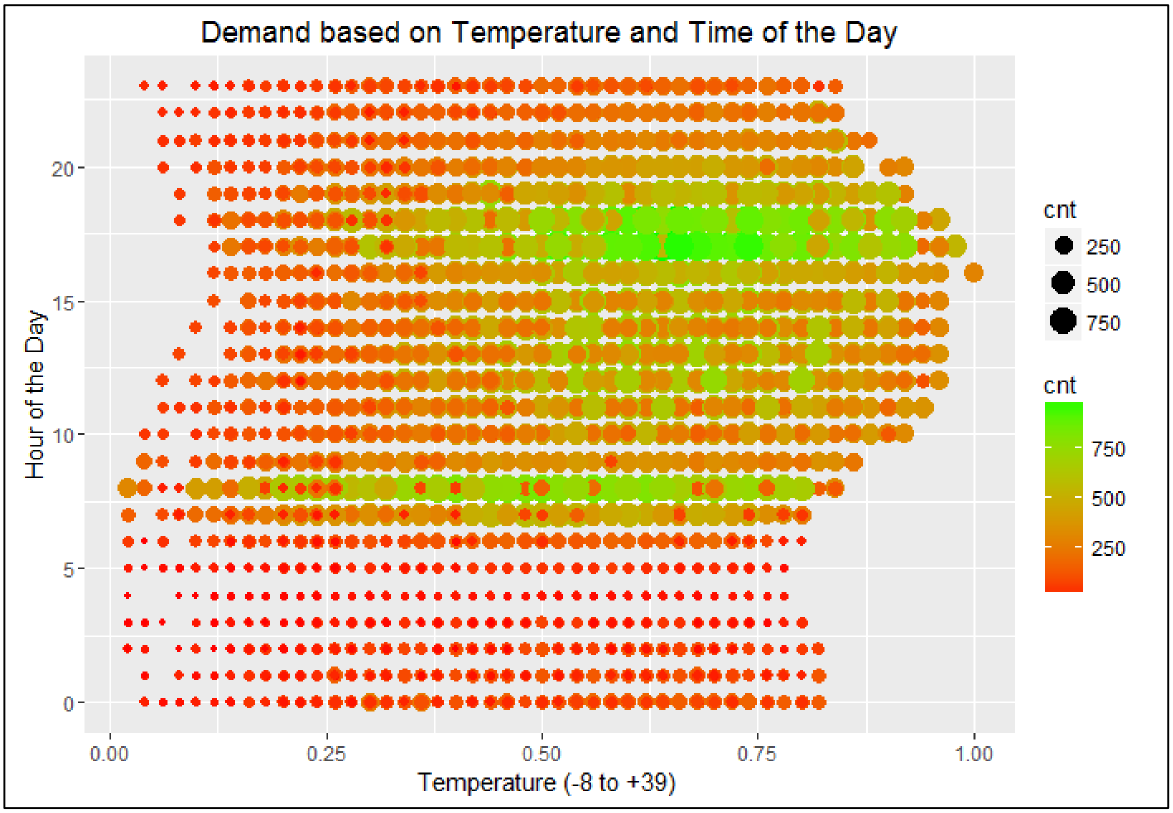
2.1.4. ANALYSIS:

The following points show various analysis made on this dataset:

a. The following diagram does an analysis of atemp variable in the data set which is based on the “feels like” temperature on weather factors. The plot below explains this relationship→

```
#Importing the data
my_data <- read.csv("BikeSharing.csv", stringsAsFactors = FALSE)

#Atemp vs. Temp
x<- ggplot(my_data, aes(temp, hr, color=cnt, size=cnt))
+ geom_point()
+    scale_colour_gradient(low="red",high="green")
+     labs(title=" Demand based on Temperature and Time of the Day")
+    labs(x="Temperature (-8 to +39")
+    labs(y="Hour of the Day")
> x
```



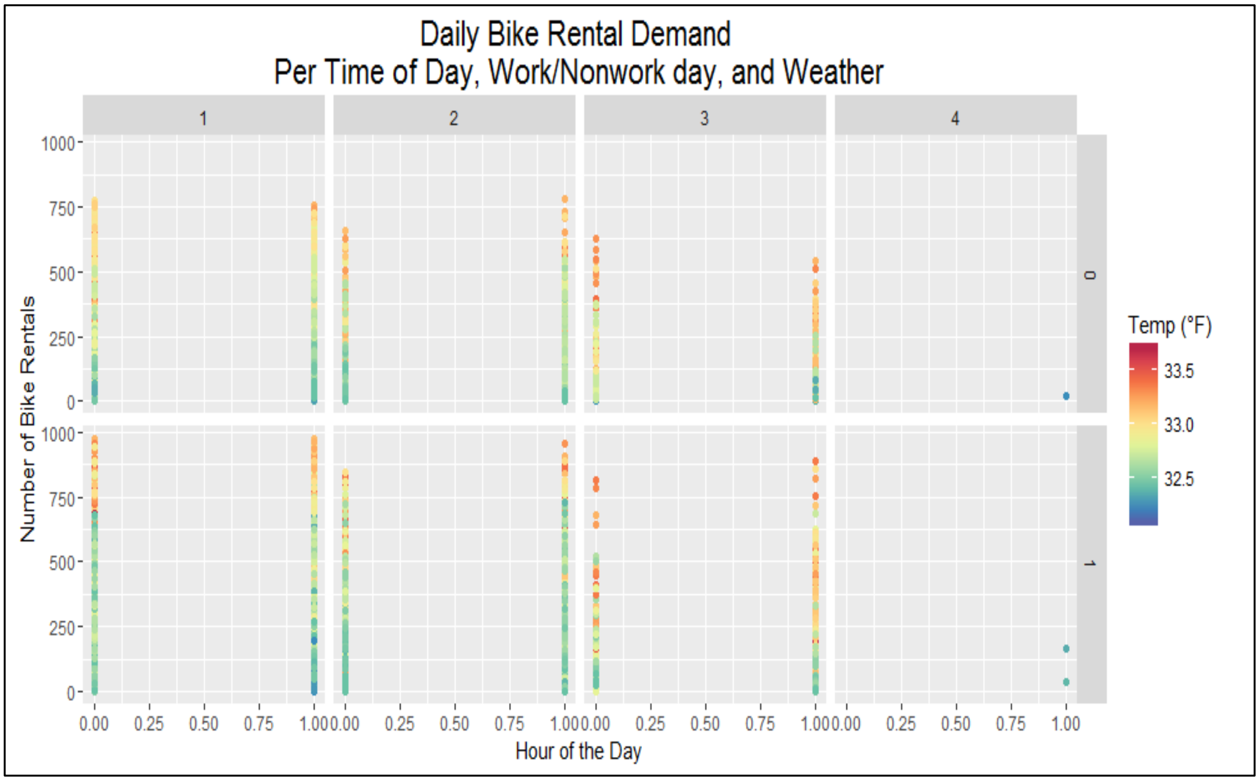
We can observe from the above diagram that:

- Demand for the bikes are heavily influenced by office peak hours.
- Temperature is directly proportional to the demand for bikes. However, when the temperature nears 35 degree C the demand is very minimal irrespective of time of the day.
- The data is skewed during morning hours whereas it is distributed in the evening.

2. The below plot shows the effect of time, temperature, weather, and work/non-work days on daily rental counts:

```
#Importing the data
> diagram2 <- ggplot(my_data, aes(my_data$workingday, cnt, color=9/5*temp+32))
+ facet_grid(workingday ~ weathersit)
+ geom_point()
+ geom_smooth()
+ theme(plot.title = element_text(size = rel(1.5)))
+ ggtitle("Daily Bike Rental Demand Per Time of Day, Work/Non-work day, and Wea
ther")
```

```
+ xlab("Hour of the Day")
+  ylab("Number of Bike Rentals")
+ scale_colour_gradientn("Temp (°F)", colours=colors.tempurature)
```

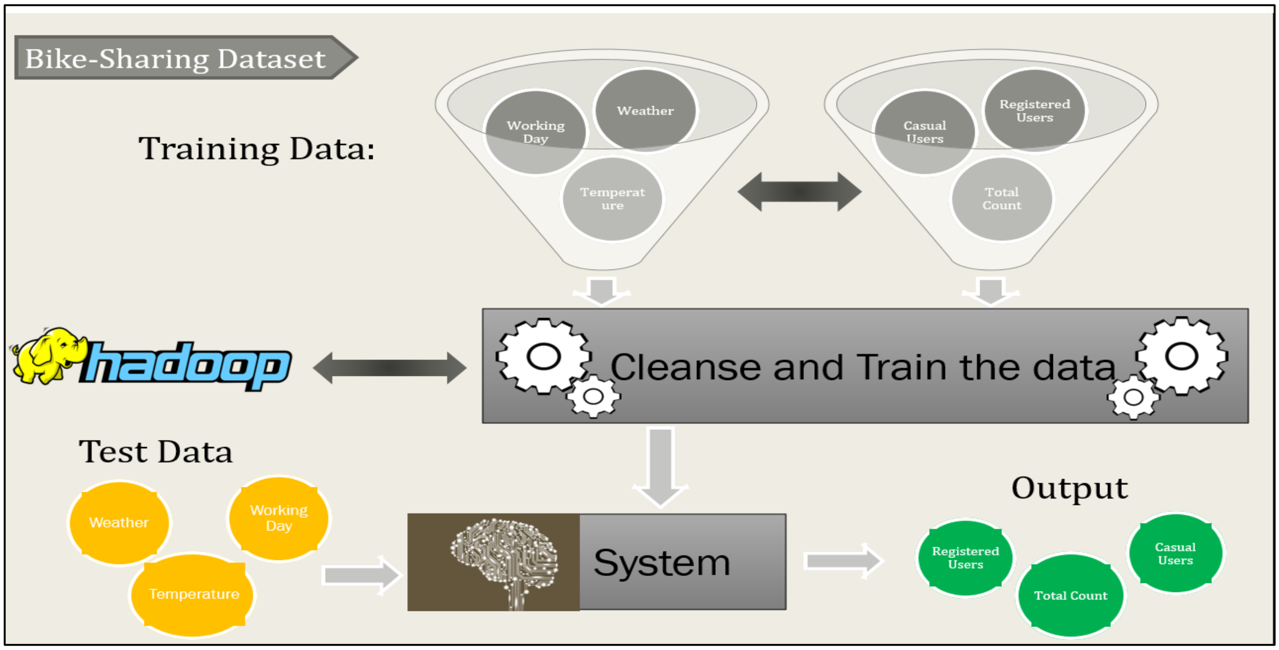


We can observe that:

- The demand for bikes is high during Clear weather or when there is a mist and broken clouds.
- The demand reduces when there is light rain and very less during heavy rains.
- There is only a marginal drop of demand on non-working days.

2.1.5. SOLUTION ARCHITECTURE & ML TECHNIQUES:

The following diagram shows the solution architecture for Bike Sharing Dataset:



We shall be taking the weather situation, working day, temperature and other suitable attributes as input data which will give us the total (registered + casual) count of users as output. We will be training this data using *Random Forest algorithm* to predict the count of users using the bike sharing system and will also implement a recommender system for the same. We will use Hadoop for data cleansing.

2.1.6. EXPECTED OUTCOME:

For a given day and hour and their respective weather attribute we shall forecast the number of casual and registered users who will using the bike sharing system

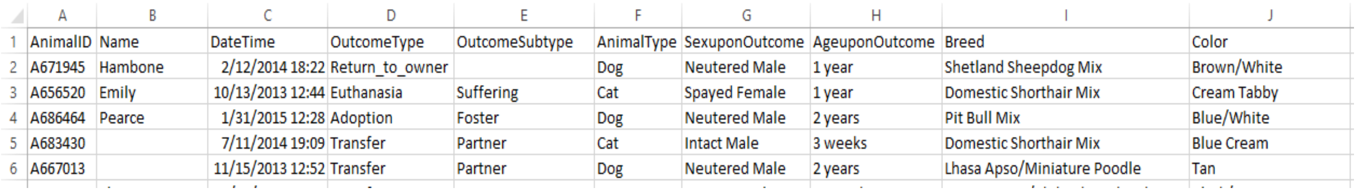
Sample Output:

dteday	season	yr	mnth	hr	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
12/31/2012	1	1	12	17	0	1	1	2	0.26	0.288	0.5	0.0896	14	150	164
12/31/2012	1	1	12	18	0	1	1	2	0.26	0.273	0.5	0.1343	10	112	122
12/31/2012	1	1	12	19	0	1	1	2	0.26	0.258	0.6	0.1642	11	108	119
12/31/2012	1	1	12	20	0	1	1	2	0.26	0.258	0.6	0.1642	8	81	89

2.2. Dataset 2→ Animal Shelter Outcomes:

The US shelter data shows that close to 30% of companion animals that comes to the adoption shelters end up being unwanted and are euthanized. The following dataset is taken from Austin Animal Shelter from October 1st, 2013 to March, 2016which we will use to predict the outcome for each animal. These insights could help shelters focus their energy on specific animals who need a little extra help finding a new home.

Following is the snapshot of the dataset:



	A	B	C	D	E	F	G	H	I	J
1	AnimalID	Name	DateTime	OutcomeType	OutcomeSubtype	AnimalType	SexuponOutcome	AgeuponOutcome	Breed	Color
2	A671945	Hambone	2/12/2014 18:22	Return_to_owner		Dog	Neutered Male	1 year	Shetland Sheepdog Mix	Brown/White
3	A656520	Emily	10/13/2013 12:44	Euthanasia	Suffering	Cat	Spayed Female	1 year	Domestic Shorthair Mix	Cream Tabby
4	A686464	Pearce	1/31/2015 12:28	Adoption	Foster	Dog	Neutered Male	2 years	Pit Bull Mix	Blue/White
5	A683430		7/11/2014 19:09	Transfer	Partner	Cat	Intact Male	3 weeks	Domestic Shorthair Mix	Blue Cream
6	A667013		11/15/2013 12:52	Transfer	Partner	Dog	Neutered Male	2 years	Lhasa Apso/Miniature Poodle	Tan

2.2.1. SUMMARY:

- *Number of Instances* = 26729
- *Number of Attributes* = 10

2.2.2. ATTRIBUTES:

- AnimalID
- Name
- DateTime
- OutcomeType: represent the status of animals as they leave the Animal Center

Adoption	Died	Euthanasia	Return to owner	Transfer

- OutcomeSubtype
- AnimalType
- SexuponOutcome
- AgeuponOutcome
- Breed
- Color

2.2.3. MACHINE LEARNING QUESTIONS:

Question 1 →
What could be the potential outcome for a new dog coming to the shelter?

Question 2 →
What could be the potential outcome for a new dog coming to the shelter?

Question 3 →
In which season/days are animals more likely to be adopted so that the shelter could make advance plans?

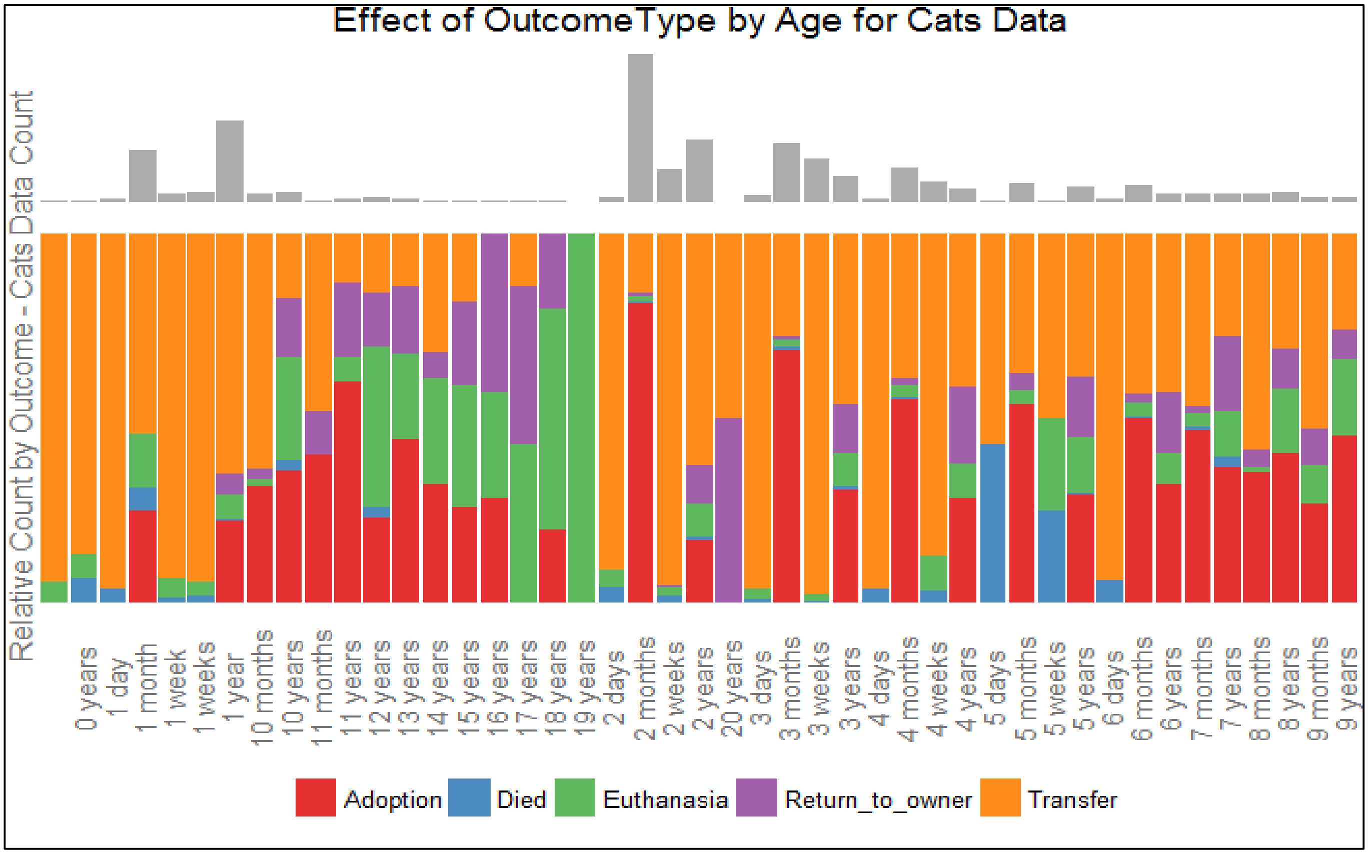
Question 4 →
For a given un-named animal (cat or a dog, gender, age, breed, sex), does giving the name increase the chances of adoption?

2.2.4. ANALYSIS:

The following points show various analysis made on this dataset:

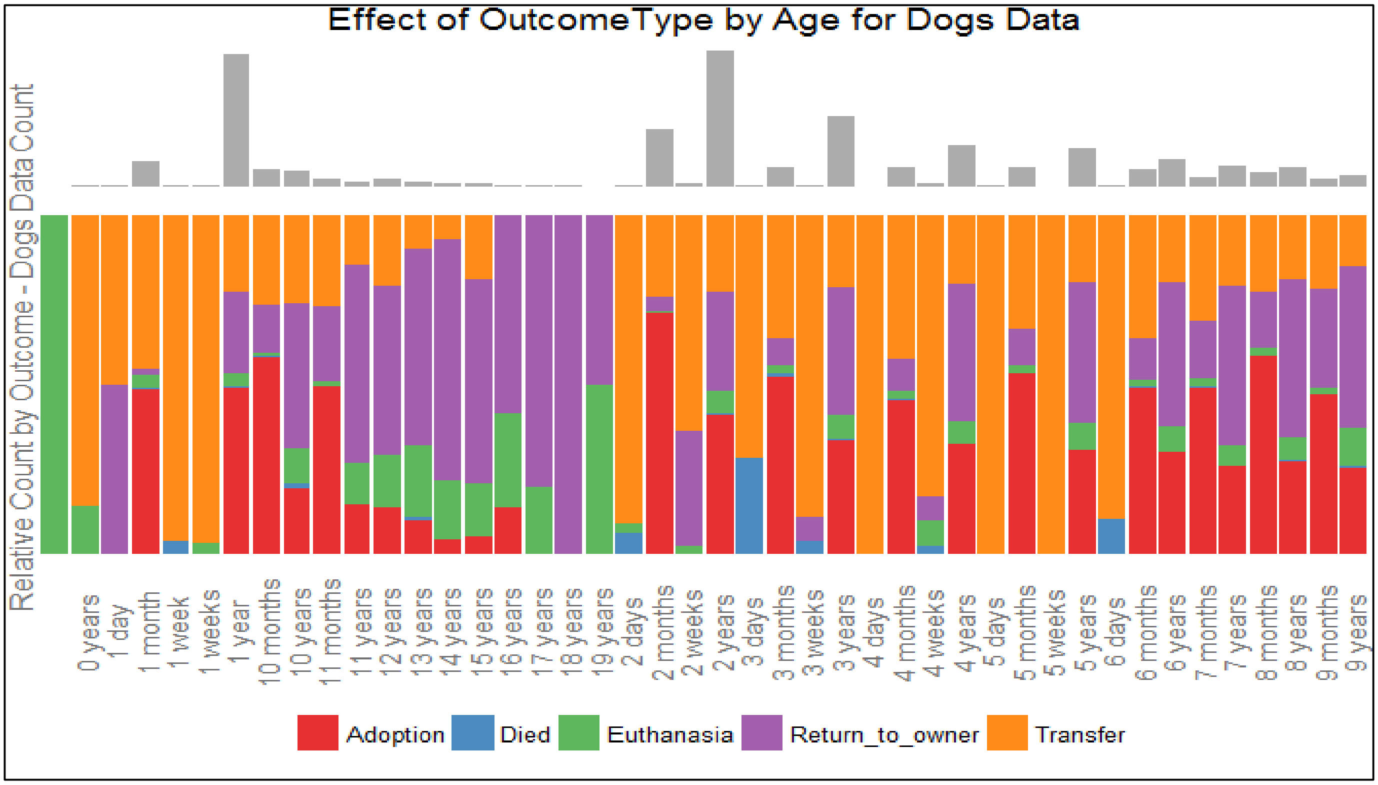
a. The following diagrams does an analysis of cats and dogs data to check what difference in puts on the outcome depending on the type of animal . The plot below explains this relationship

```
#Filtering cat data from the dataset
> cat_data <- filter(Animal_data, AnimalType == "Cat")
#How does the OutcomeType affect the age upon Outcome
> cat_data %>% count(AgeuponOutcome) %>%
+ ggplot(aes(x = AgeuponOutcome, y = n))
+ geom_bar(stat = "identity", alpha = 0.5)
+ theme_void()
+ ylab("Count")
+ggtitle("Effect of OutcomeType by Age for Cats Data")
+ theme(axis.title.y = element_text(angle = 90, color = "#737373" )) -> g1
# Relative Count by Outcome – Cats Data
> ggplot(cat_data, aes(x = AgeuponOutcome, fill = OutcomeType))
+ geom_bar(stat = "count", position = "fill",  alpha = 0.9)
+ scale_fill_brewer(palette = "Set1")
+ theme_void()
+ ylab("Relative Count by Outcome") +
+ theme(axis.text.x = element_text(angle = 90, vjust = 0.5, size = 11, color = "#737373"),
      legend.position = "bottom",
        axis.title.y = element_text(angle = 90, color = "#737373")) -> g2
# To bring up the above 2 Graphs on the same page
> grid.arrange(g1, g2, ncol = 1, heights = c(1,3))
```



Similar data retrieved for dogs:

```
#Filtering cat data from the dataset
> dog_data <- filter(Animal_data, AnimalType == "Dog")
#How does the OutcomeType affect the age upon Outcome
> dog_data %>% count(AgeuponOutcome) %>%
+ ggplot(aes(x = AgeuponOutcome, y = n))
+ geom_bar(stat = "identity", alpha = 0.5)
+ theme_void()
+ ylab("Count")
+ggtitle("Effect of OutcomeType by Age for dogs data")
+ theme(axis.title.y = element_text(angle = 90, color = "#737373" )) -> g3
# Relative Count by Outcome – Dogs Data
> ggplot(dog_data, aes(x = AgeuponOutcome, fill = OutcomeType))
+ geom_bar(stat = "count", position = "fill",  alpha = 0.9)
+ scale_fill_brewer(palette = "Set1")
+ theme_void()
+ ylab("Relative Count by Outcome") +
+ theme(axis.text.x = element_text(angle = 90, vjust = 0.5, size = 11, color = "#737373"),
        legend.position = "bottom",
        axis.title.y = element_text(angle = 90, color = "#737373")) -> g4
# To bring up the above 2 Graphs on the same page
> grid.arrange(g3, g4, ncol = 1, heights = c(1,3))
```

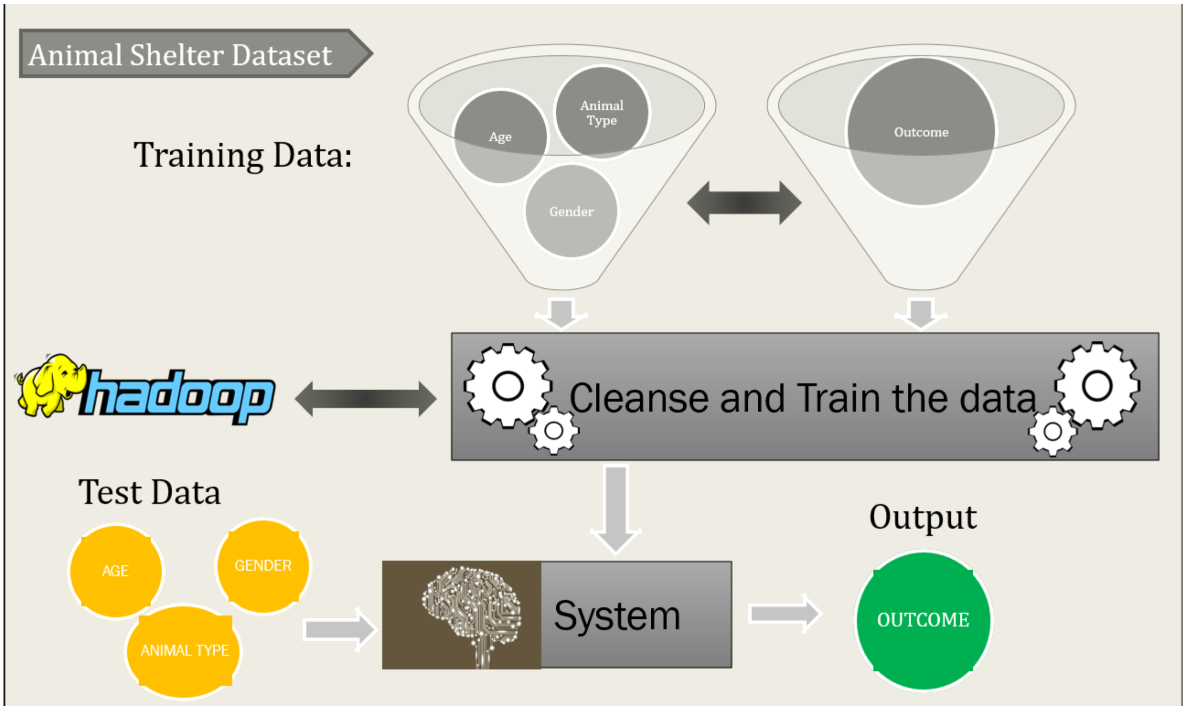


We observe from the above 2 graphs:

- For Cats: The adoption rate is relatively high for older cats
- For Dogs: The adoption rate is very high for young dogs and puppies. As the age increases the adoption rate falls.
- Older dogs have higher chances of getting returned to owner.

2.2.5. SOLUTION ARCHITECTURE & ML TECHNIQUES:

The following diagram shows the solution architecture for Animal Shelter Outcome Dataset:

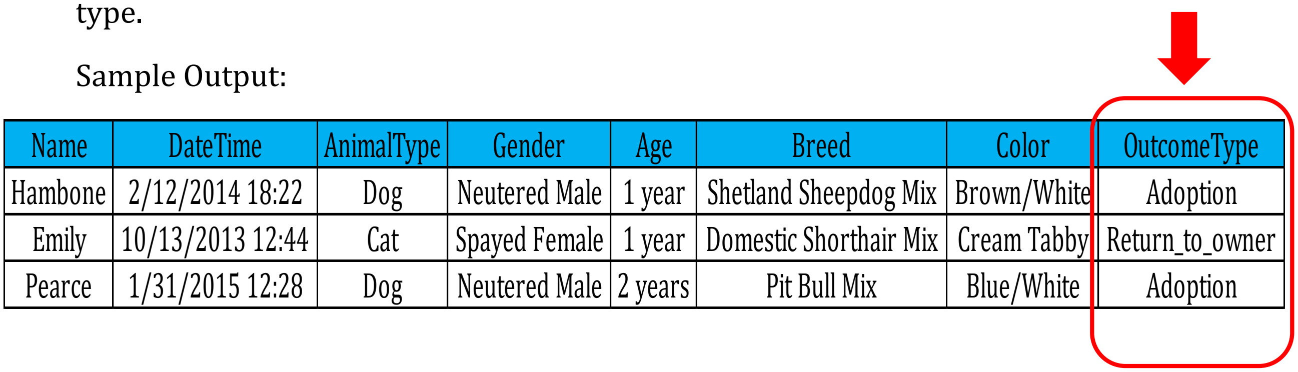


We shall be taking the animal type, age, gender, breed and other suitable attributes as input data which will give us the outcome as output. We will be training this data using *Random Forest algorithm* to predict the outcome and will also implement a recommender system to suggest names for unnamed animals which will increase their chances for adoption. We will use Hadoop for data cleansing.

2.2.6. EXPECTED OUTCOME:

When a new dog or a cat comes to the shelter we should be able to predict the outcome type.

Sample Output:



Name	DateTime	AnimalType	Gender	Age	Breed	Color	OutcomeType
Hambone	2/12/2014 18:22	Dog	Neutered Male	1 year	Shetland Sheepdog Mix	Brown/White	Adoption
Emily	10/13/2013 12:44	Cat	Spayed Female	1 year	Domestic Shorthair Mix	Cream Tabby	Return_to_owner
Pearce	1/31/2015 12:28	Dog	Neutered Male	2 years	Pit Bull Mix	Blue/White	Adoption

3. References:

1. https://cran.r project.org/web/packages/gridExtra/vignettes/arrangeGrob.html
2. https://cran.r-project.org/web/packages/gridExtra/gridExtra.pdf
3. http://www.sthda.com/english/wiki/ggplot2-easy-way-to-mix-multiple-graphs-on-the-same-page-r-software-and-data-visualization
4. https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf
5. Bike Sharing Dataset: http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset
6. Animal Shelter Outcome Dataset: https://www.kaggle.com/c/shelter-animal-outcomes/data