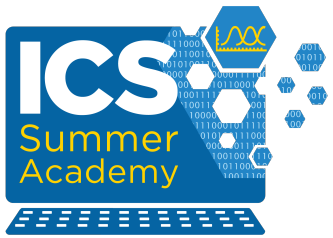
AIzheimer's Data Analysis

Babak Shahbaba and Sam Behseta
UC Irvine and California State University Fullerton

July and August, 2022



Let's Begin!

- In the second part of morning sessions, we apply our knowledge of data science methods and techniques in order to detect and better understand the contributing factors to Alzheimer's disease, through analyzing the National Alzheimer's Coordinating Center's (NACC).

Let's Begin!

- In the second part of morning sessions, we apply our knowledge of data science methods and techniques in order to detect and better understand the contributing factors to Alzheimer's disease, through analyzing the National Alzheimer's Coordinating Center's (NACC).
- This data set is cleaned and is only a small subset of a significantly larger data set.

Let's Begin!

- In the second part of morning sessions, we apply our knowledge of data science methods and techniques in order to detect and better understand the contributing factors to Alzheimer's disease, through analyzing the National Alzheimer's Coordinating Center's (NACC).
- This data set is cleaned and is only a small subset of a significantly larger data set.
- Nevertheless, as far as the objectives of our course matters, it helps us to understand some of the factors contributing to dementia.

Let's Begin!

- One last point before we begin: the original data set came with a significant number of missing values. A number of graduate students at UCI have worked on *imputing* those missing values. Missing values imputation is a set of sophisticated statistical methods for estimating data points that are not available to us, mainly due to random mechanisms.

A Quick Tour of the NACC Data Set

- What is the size of this data? Remember that rows are patients, and columns are *features* or *variables* associated with each patient.

A Quick Tour of the NACC Data Set

- What is the size of this data? Remember that rows are patients, and columns are *features* or *variables* associated with each patient.
- I would like to start by giving my working data set an easy title in R, and quickly figure out its dimensions. So, let's start by some coding:

```
alz<-alzheimer_data
dim(alz)
```

A Quick Tour of the NACC Data Set

- What is the size of this data? Remember that rows are patients, and columns are *features* or *variables* associated with each patient.
- I would like to start by giving my working data set an easy title in R, and quickly figure out its dimensions. So, let's start by some coding:

```
alz<-alzheimer_data
dim(alz)
```

- OK! so, we have 2700 patients and 57 features for each patient. My next move is to set the now called *alz* data set, as my default or working data set.

```
attach(alz)
length(hrate)
```

A Quick Tour of the NACC Dataset

- As you recall from Yueqi's introduction, some of our variables are categorical and some are numerical. Moreover, they have different range and label values. It is useful to get a quick snapshot od the nature of the features in our data set:

```
str(alz)
```

A Quick Tour of the NACC Dataset

- As you recall from Yueqi's introduction, some of our variables are categorical and some are numerical. Moreover, they have different range and label values. It is useful to get a quick snapshot od the nature of the features in our data set:

```
str(alz)
```

- You can always ask for R to show you the first few rows (standard is 6) of the entire data set:

```
head(alz)
```

A Quick Tour of the NACC Dataset

- As you recall from Yueqi's introduction, some of our variables are categorical and some are numerical. Moreover, they have different range and label values. It is useful to get a quick snapshot od the nature of the features in our data set:

```
str(alz)
```

- You can always ask for R to show you the first few rows (standard is 6) of the entire data set:

```
head(alz)
```

- Or ask R to show you a specific part of data set:

```
alz[2,]
alz[3:4,1:5]
```

A Quick Tour of the NACC Dataset

- Along those lines, you can create smaller data set for specific goals. See the code below, as an example:

```
alz.test=alz[,c("id","diagnosis","educ")]
head(alz.test)
```

- Or manage the same thing via *tidyverse* package:

```
alz.test=select(alz,id,diagnosis,educ)
head(alz.test)
```

A feature called *diagnosis*

- We can then focus on certain variables of interest. Suppose we would like to learn more about the feature tagged as *diagnosis*. Recall that this feature has three outcomes, 0= normal cognitive diagnosis, 1= mild demential symptoms, and 2= strong symptoms of dementia. 0, 1, and 2 are merely levels or *factors* of this feature. As such, we can tell R to treat this variable as a *factor* variable.

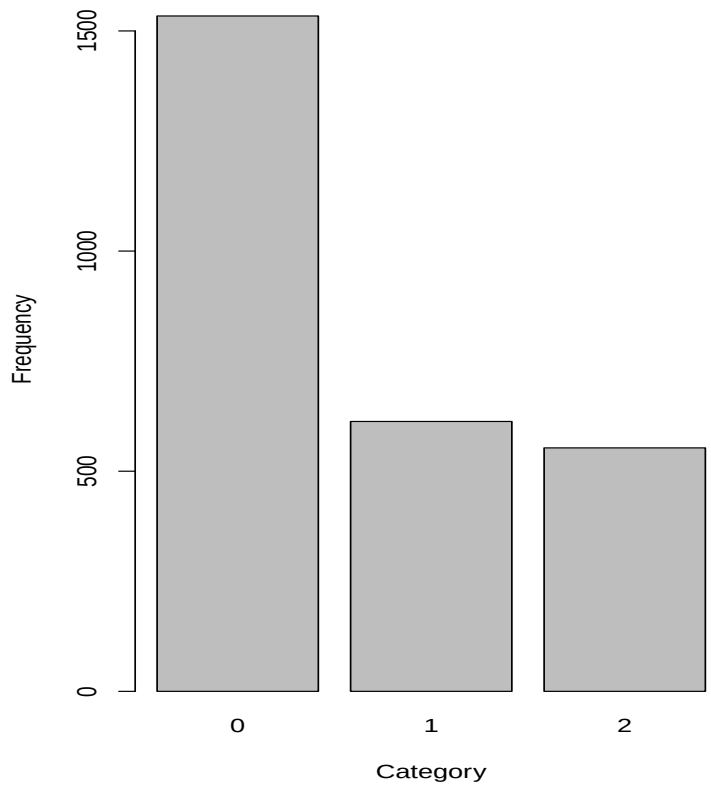
```
diag<-as.factor(diagnosis)
summary(diag)
```

```
> summary(diag)
   0    1    2
1534  613  553
```

Visualizing Diagnosis

- Barplot is the basic tool for visualizing categorical variables.

```
plot(diag,xlab="Category", ylab="Frequency",
ylim=c(0,1600))
```

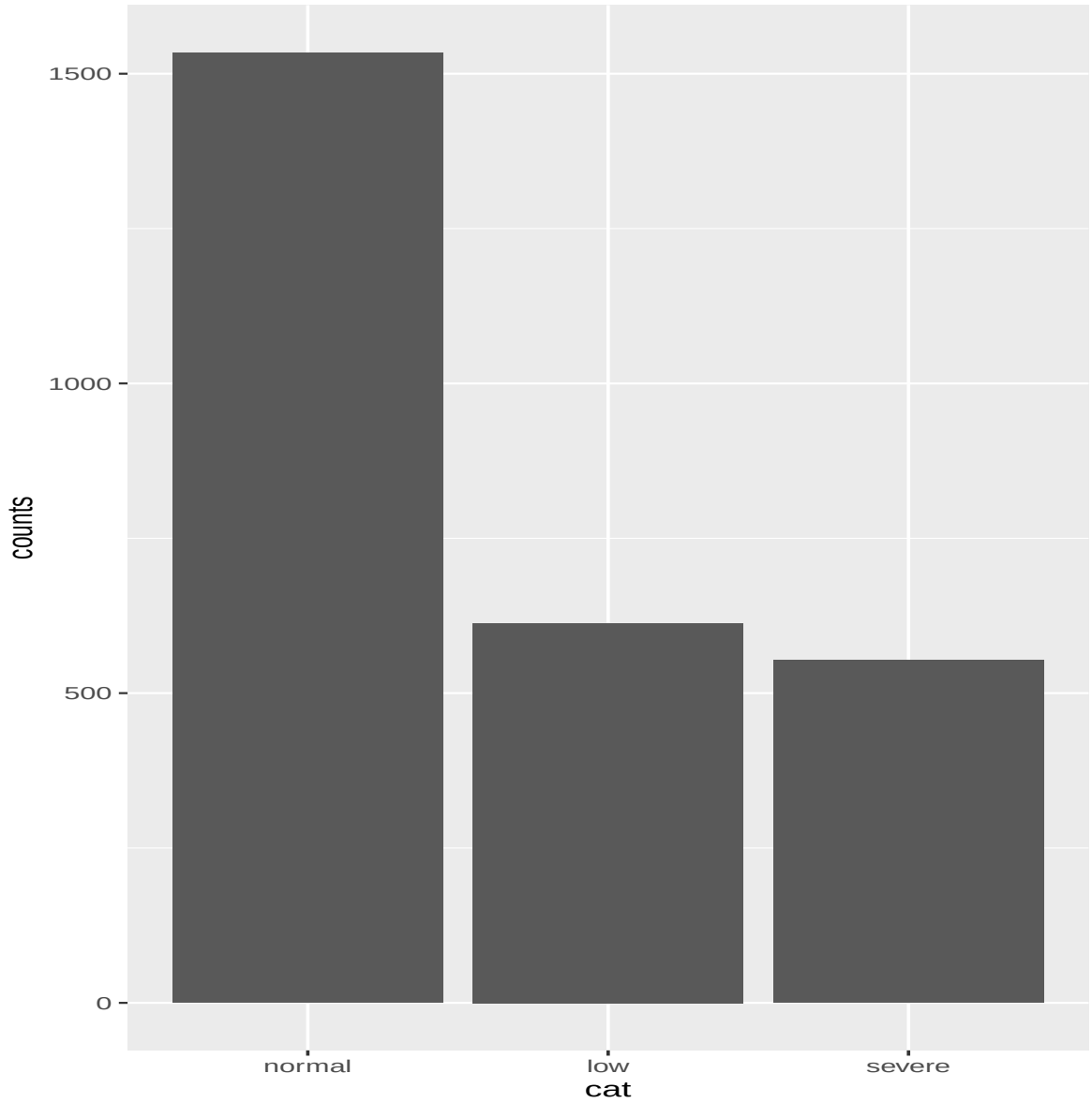


Visualizing Diagnosis

- You can always improve the quality of your visuals in R with the *ggplot2* package.

```
library(ggplot2)
counts<-c(1534,613,553)
cat<-c("normal","low","severe")
cat=as.factor(cat)
df<-data.frame(counts,cat)
p<-ggplot(data=df, aes(x=cat, y=counts)) +
  geom_bar(stat="identity")
p +
scale_x_discrete(limits=c("normal", "low", "severe"))
```

Visualizing Diagnosis



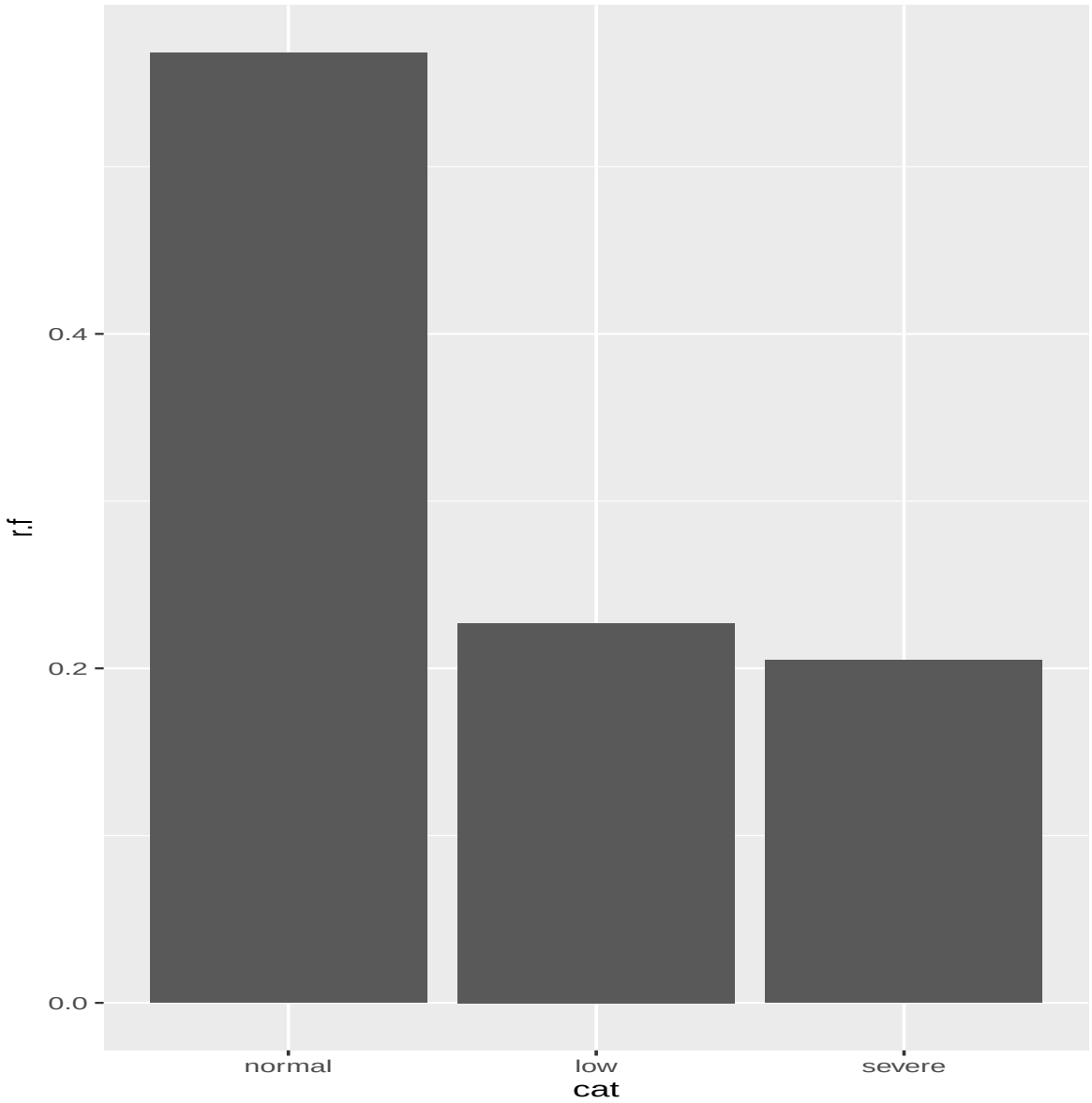
Visualizing Diagnosis

- It makes sense to work with *relative frequencies* as opposed to the frequencies.

```
r.f=counts/sum(counts)

df=data.frame(r.f,cat)
p<-ggplot(data=df, aes(x=cat, y=r.f)) +
  geom_bar(stat="identity")
p +
scale_x_discrete(limits=c("normal", "low", "severe"))
```

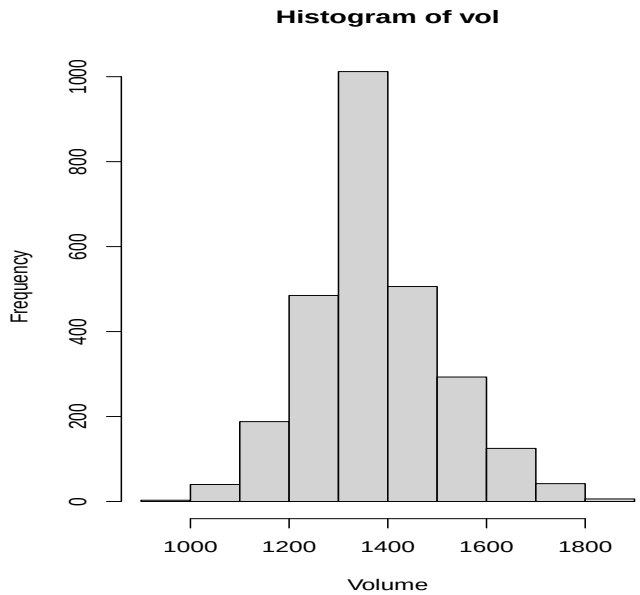
Visualizing Diagnosis



Visualizing Total Intracarnial Volume

- As Alzheimer's progresses, the brain volume can significantly decrease. In other words, the cortex overall becomes thinner or the brain gradually shrinks. For a feature representing brain volume, we can focus on *NACCICV*. Note that this is a numerical feature.

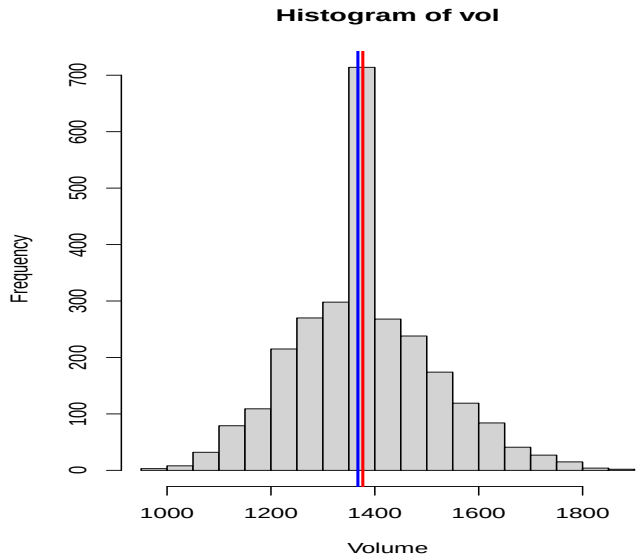
```
vol<- naccicv
summary(vol)
hist(vol,xlab="Volume")
```



Visualizing Total Intracarnial Volume with a Histogram

- This looks like a *symmetric* distribution (more on that later!) Note how the mean and the median are so close to each other. We can make the histogram finer: while the shape of the distribution does not change much, it reveals a large number of patients with the brain volume around 1380-1400!

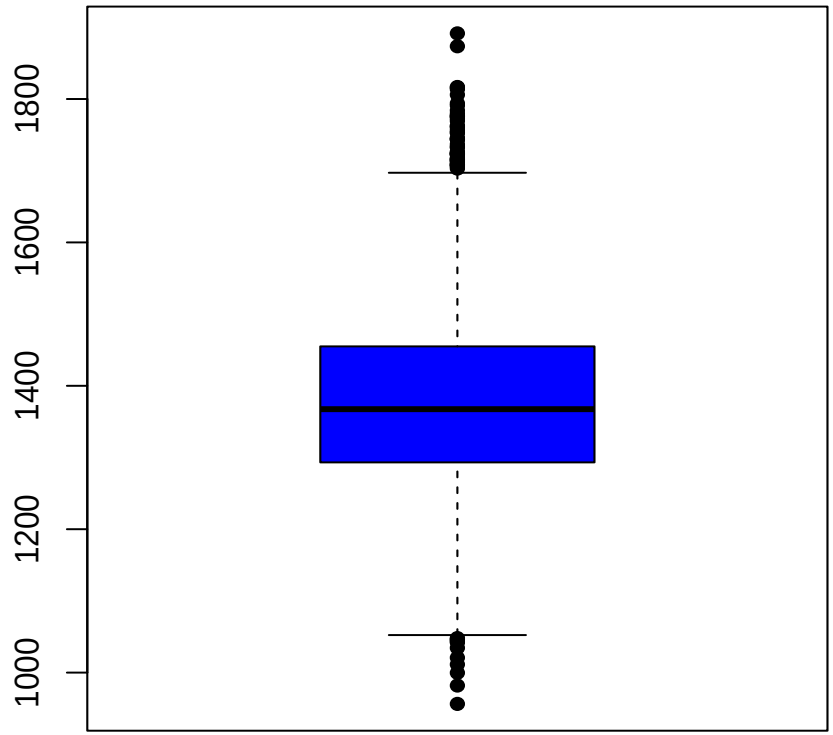
```
vol<- NACCICV
summary(vol)
hist(vol,xlab="Volume",breaks=15)
abline(v=mean(vol),col="red",lwd=2)
abline(v=median(vol),col="blue",lwd=2)
```



Visualizing Total Intracarnial Volume with a Boxplot

- In base R, we can create boxplots using the command called, well *boxplot*!

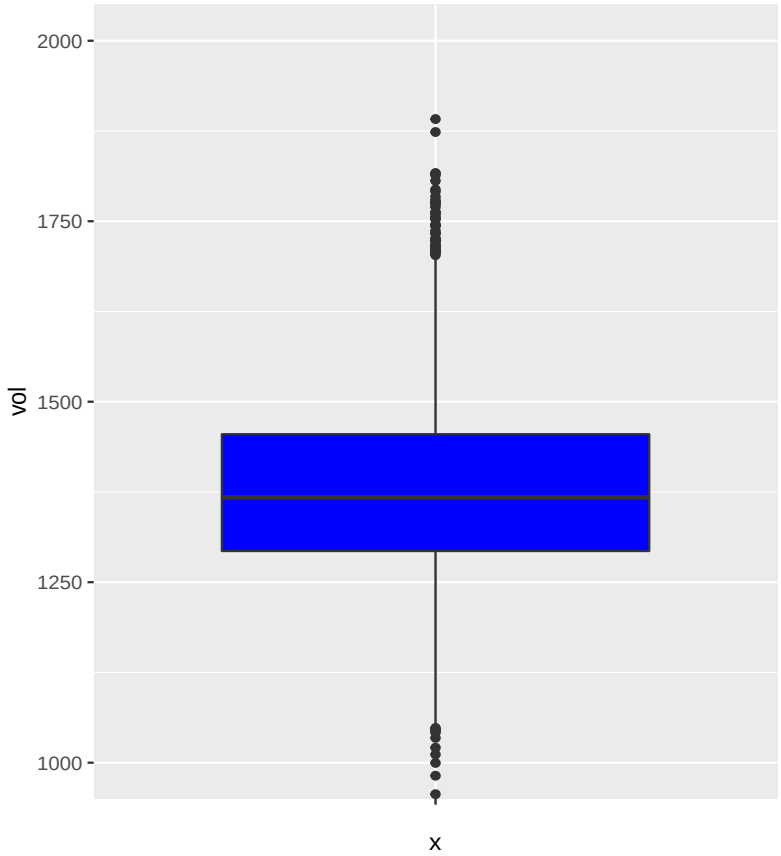
```
boxplot(vol,col ="blue",pch=16)
```



Visualizing Total Intracarnial Volume with a Boxplot

- Alternatively we can use ggplot2 for creating the same boxplot.

```
vol1=data.frame(vol)
ggplot(data = vol1, aes(x = "", y = vol)) +
  geom_boxplot(fill="blue") +
  coord_cartesian(ylim = c(1000, 2000))
```



Back to Boxplot

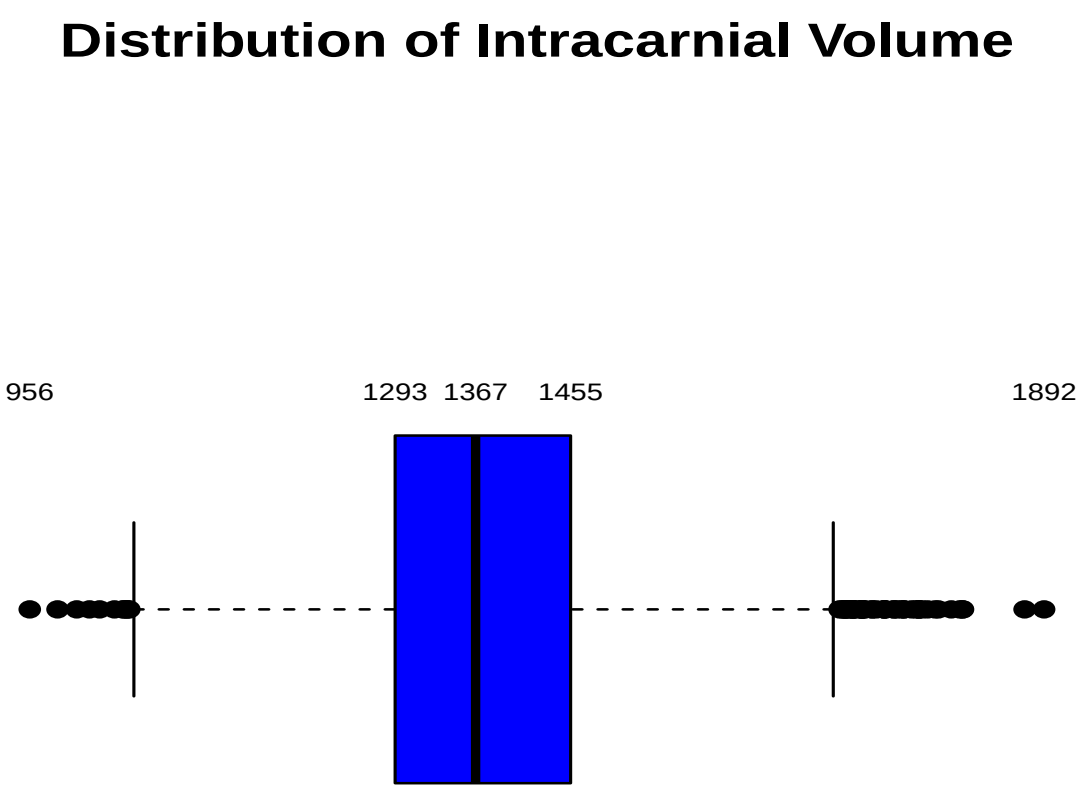
- Let's identify and highlight the important characteristics of the boxplot we just made.

```
q.vol=quantile(vol)
q.vol
range(vol)
inter.quartile=IQR(vol)
outlier.lower.threshold=q.vol[2]-1.5*inter.quartile
outlier.upper.threshold=q.vol[4]+1.5*inter.quartile
thresholds=c(outlier.lower.threshold,outlier.upper.threshold)
thresholds
```

Back to Boxplot

- Now let's use the information we gained and add it to the boxplot!

```
boxplot(vol,col ="blue",horizontal=TRUE,axes=FALSE,
        pch=16)
text(x=fivenum(vol), labels=round(fivenum(vol)),
     y=1.25,cex=0.6)
```



Class Activity 1 – Group Work

- Analyze the variables *female* and *csfvol* (research it!) Summarize each feature and provide appropriate visuals. Remember to label the axes!
- Take a logarithm of *csfvol* (research it!) and analyze the transformed variable! Any significant changes?
- Write a paragraph on each of the features. Make sure you discuss the shape of distributions and whether there are outliers.
- Exchange your work with the nearest team around you. Discuss and compare their work and grade their write ups (on a scale of 0 to 100)!

Lab 1 – Group Work on Thursday

- Identify all categorical and numerical variables in the data set. Use the data dictionary, coupled with your own research to understand what each variable represents and its potential role in studying Alzheimer's disease.
- Give a thorough analysis (summarization and visualization) of all the numerical features and at least 2 categorical variables.