Exploratory data analysis

Objective

- Get the quick idea about data
  - visualization is the easiest way
  - check descriptive statistics
- Data cleaning process to reduce the number of data problems in the future
  - handle missing data, outliers or typo etc.
  - need to be careful!
- Explore your data to determine whether the model assumptions are met etc.
  - E.g., check normality of data

Visualization and descriptive statistics

- **Visualization**
  - Histogram
  - Boxplot
  - Scatter plot (to find the relationship btw 2 variables)

- **Descriptive statistics**
  - Mean, median, variance, skewness, kurtosis etc.
  - Correlation (for 2 variables)

- Get rough idea about the distribution of data
- Check outliers or missingness
- In general, to check normality of data

Visualization and descriptive statistics

- Skewed right: mean > median
- Skewed left: mean < median
  - **Robustness** of median.
  - Able to guess its skewness based on mean an median values

- Able to check its normality (informally) based on visual and descriptive statistics

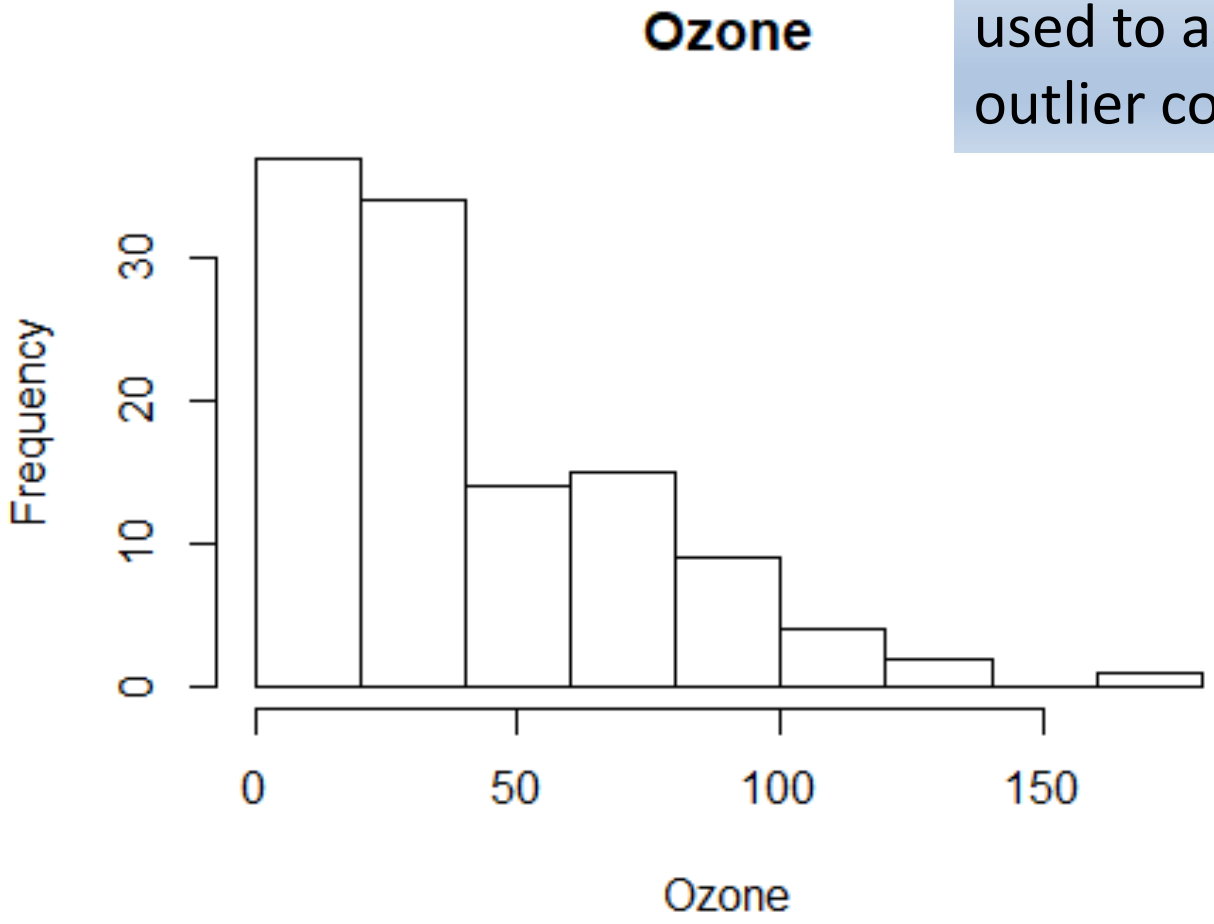
Example : airquality

- Daily air quality measurements in New York, May to September 1973. (R built-in data)
- 154 observations on 6 variables – Ozone, Solar R, Wind, …

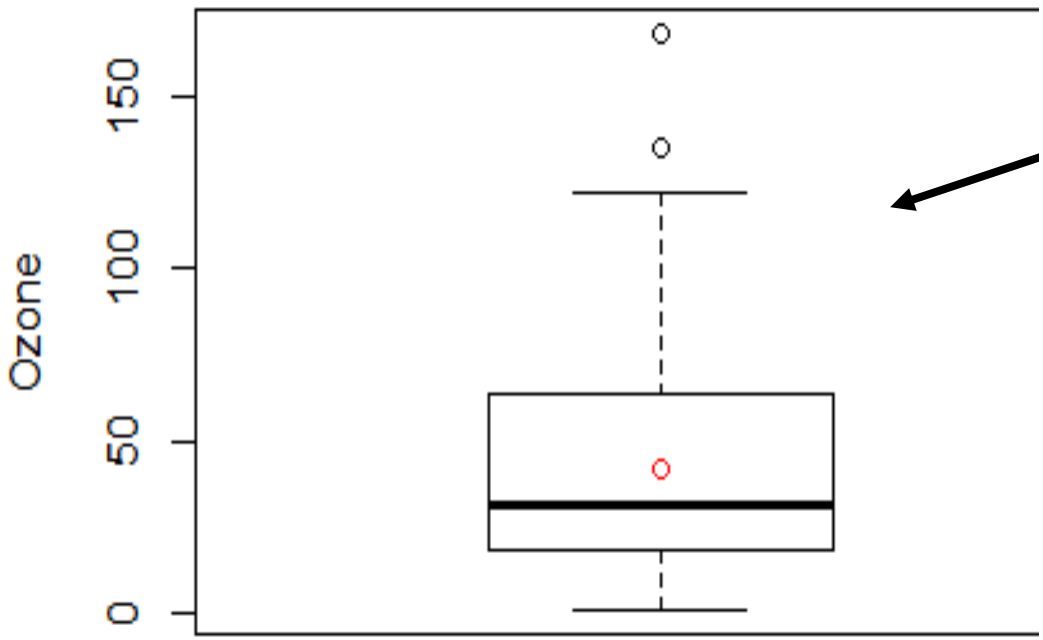
```
> head(airquality,10)
   Ozone Solar.R Wind Temp Month Day
1     41     190  7.4   67     5   1
2     36     118  8.0   72     5   2
3     12     149 12.6   74     5   3
4     18     313 11.5   62     5   4
5     NA      NA 14.3   56     5   5
6     28      NA 14.9   66     5   6
7     23     299  8.6   65     5   7
8     19      99 13.8   59     5   8
9      8      19 20.1   61     5   9
10    NA     194  8.6   69     5  10
>
```

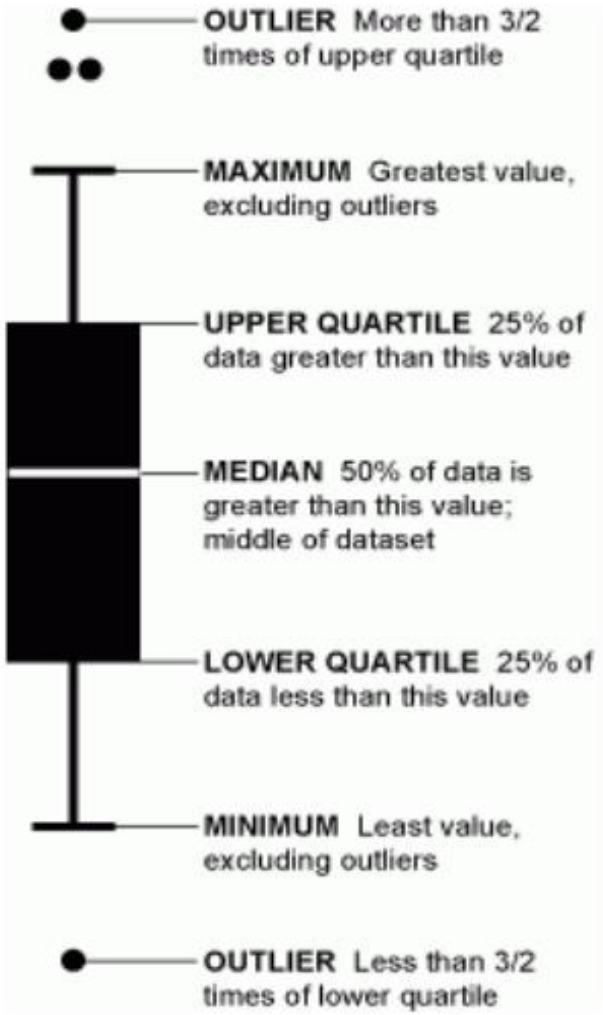
```
hist(airquality$Ozone,main="Ozone",xlab="Ozone")
```

Provides the distribution of the data. This can also be used to assess potential outlier concerns.



```r
boxplot(airquality$Ozone,ylab="Ozone")
points(mean(airquality$Ozone, na.rm=TRUE), col="red")
```





Example of descriptive statistics

```
summary(airquality$Ozone, na.rm=TRUE)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.      NA's
##    1.00   18.00   31.50   42.13   63.25  168.00        37
mean(airquality$Ozone, na.rm=TRUE) ->  mean
## [1] 42.12931
var(airquality$Ozone, na.rm=TRUE) -> variance
## [1] 1088.201
skewness(airquality$Ozone, na.rm=TRUE) -> skewness
## [1] 1.209866
range(airquality$Ozone,na.rm=TRUE) -> range [min,max]
## [1]   1 168
```

Missing data handling

- Can be a separate semester-long course
- Missing mechanisms:
  - Missing Completely at Random (MCAR)
    - Missing occurs by random
  - Missing at Random (MAR)
  - Missing Not at Random (MNAR)

Important statistical assumptions

- Normality
  - Why normality check is important?

    1) When conducting a **t-test** or **ANOVA,** normality assumption is required

    2) When using **correlation** and **regression techniques**, lack of normality and outliers impact your conclusions

- Normal distribution is symmetric, bell-shaped
  - Inverse is NOT true e.g., Cauchy distribution, t-distribution

  - There are a lot of tests one can use to check for normality and outliers in the data.

Inference based on Normality

- Under normality assumption, we can perform following tests.
  - ✓ One-sample t-test

    (e.g., test if iphone battery life span > 2 years)
  - ✓ Two-sample t-test

    (e.g., test if iphone and galaxy have the same life span)
  - ✓ ANOVA test (simply speaking, comparing group means among more than two groups)

    (e.g., test among iphone, galaxy and Android phone)

Detection of Normality

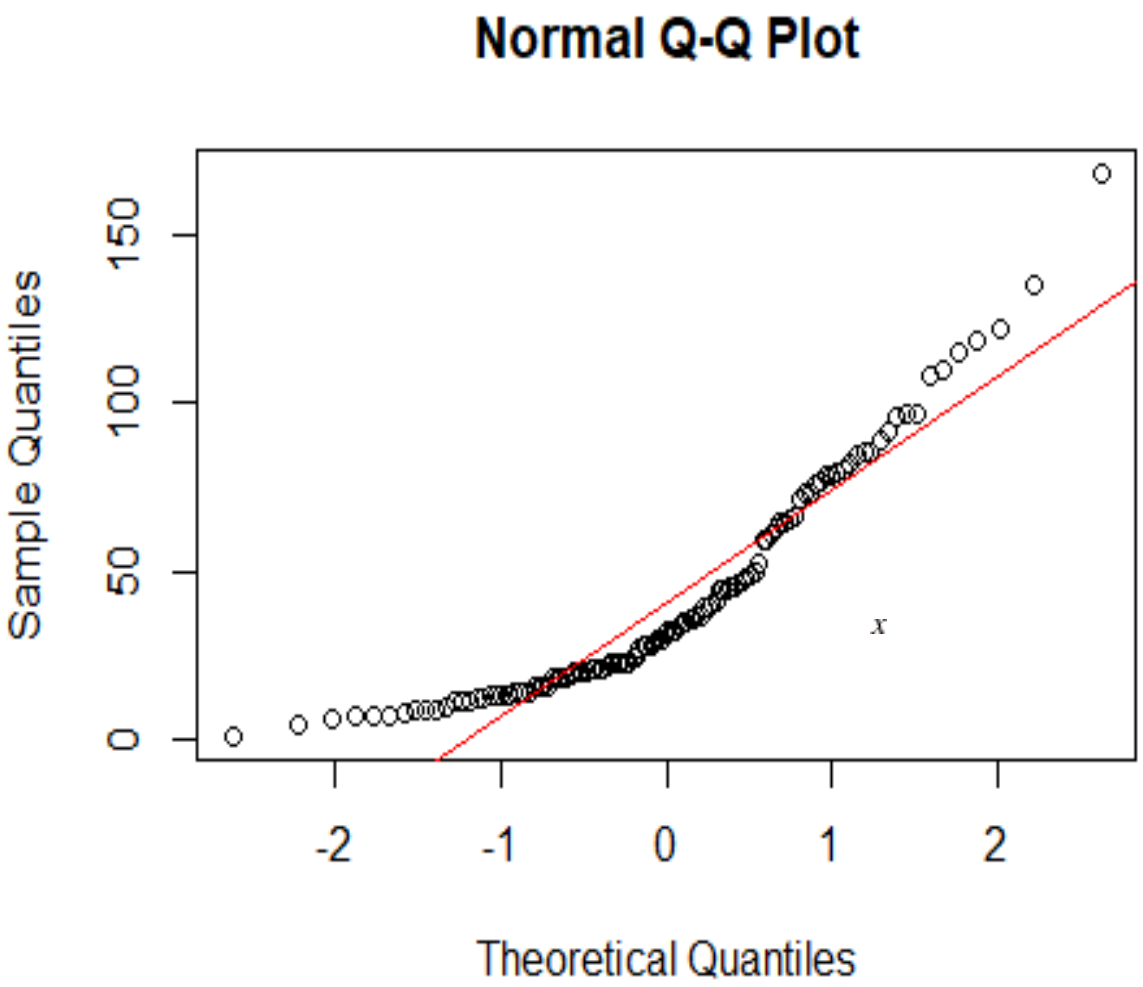
- How to check Normality?
  - ✓ **Qualitatively check** by looking at:

    : histogram, boxplot, quantile-quantile plot (QQ plot) etc..
  - ✓ **Quantitative check** by formal test

    : Sharpiro-Wilk test …

- For a comparison among groups (e.g., t-test, ANOVA), **normality check should be conducted by groups**
- If **at least one group** does not follow normality, t-test or ANOVA conclusions may **NOT** be valid.

```
qqnorm(airquality$Ozone); qqline(airquality$Ozone, col = 2)
```



Quantile-Quantile Plots (a.k.a., Q-Q plots): A useful diagnostics of how well a specified theoretical distribution fits your data. If the quantiles of the theoretical and data distributions agree, the plotted points fall on or near the line.

Shapiro-Wilk Normality test

```
shapiro.test(airquality$Ozone)

##
##  Shapiro-Wilk normality test
##
## data:  airquality$Ozone
## W = 0.87867, p-value = 2.79e-08
```

H0: Data follows normal distribution

H1: Data does not follow normal distribution

- If p-value is larger than significance level (in general α=0.05), we do not enough evidence to reject the null hypothesis, thus our conclusion is - data follows normal distribution
- If p-value is smaller than significance level, we have enough evidence to reject the null hypothesis, thus our conclusion is – data does not follow Normal distribution