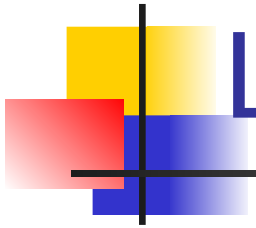
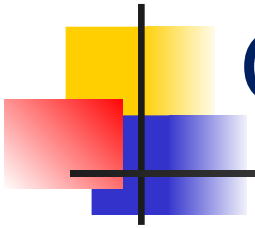


Lecture 10

- Exploratory Data Analysis
- Questions
- Variations
- Data Visualization
- Histogram, Scatterplot, Boxplot

Exploratory Data Analysis

- Exploratory Data Analysis (EDA) is the process of analyzing and visualizing data to get a better understanding of the data and obtaining insight from it. Steps involved includes:
  - Importing the data
  - Cleaning the data
  - Processing the data
  - Visualizing the data
- Exploratory Data Analysis is the critical process of doing initial investigations on data in order to:
  - Discover patterns in the data
  - Uncover underlying structure in the data
  - Detect outliers and anomalies in the data
  - Test underlying assumptions about the data
  - Extract important variables from the data
  - Determine optimal factor settings from the data
- Tools used for Exploratory Data Analysis
  - Summary Statistics
  - Graphical Representation



Questions

- “There are no routine statistical questions, only questionable statistical routines.”

  — Sir David Cox

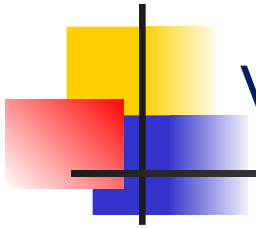
- “Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.”

  — John Tukey

- Two types of questions will always be useful for making discoveries within your data.
  - What type of variation occurs within the variables?
  - What type of covariation occurs between the variables?

Some Keywords

- Variable is a quantity, quality, or property that you can measure.
- Value is the state of a variable when is measured, which may change from measurement to measurement.
- Observation is a set of measurements made under similar conditions. An observation contains several values, each associated with a different variable.
- Tabular Data is the tendency of the values of a variable to change from measurement to measurement.
- Variation is the tendency of the values of a variable to change from measurement to measurement.
- Descriptive Data are brief informational coefficients that summarize a given data set, which represents the entire population or a sample of the population. Can be a:
  - measure of central tendency
  - measure of variability (spread)



Variations

In some situations, we may want to compare the distributions of two (or more) data sets.

A common scenario is when you need to sample from a larger dataset in order to reduce the time required for hyper-parameters tuning. In this case, you want the sample follows the same distribution of the original dataset (population) to guarantee the tuned model works properly in production.

Another possible scenario is when you want to compare the distribution between your training and testing sets to confirm the required level of similarity between them.

In both cases the problem arises when you need to compare two samples with multiple predictors (features).

There are several ways to compare populations using sample data. These can be grouped into two categories, depending on whether the population density is known or unknown.

- Visualization Approach
- Statistical Approach

Data Visualization

- Data presentation is the process of using graphical formats to visually represent two or more data sets in order that an informed decision can be made based on them.

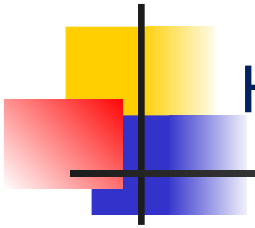
- A quick look may suggest the following:
  - a possible probability model to use
  - suitable statistical methods for the given data
  - presence or absence of outliers
  - presence or absence of heterogeneity
  - existence of time trends or other patterns
  - relation between two or several variables
- There are several graphical data representation Methods:

  e.g., Histogram, Scatter Plot, Box Plot, etc.

- Sample Data:

  To evaluate the effectiveness of a processor for a certain task, the CPU time for $n = 30$ randomly chosen jobs (in seconds) is recorded as:

70	36	43	69	82	48	34	62	35	15	59	139	46	37	42
30	55	56	36	82	38	89	54	25	35	24	22	9	56	19



Histogram

- Histogram shows the shape of the density of the data, can be used to show presence of homogeneity, and suggests possible outliers.

- The range of the data is divided into equal bins and the number of samples in each bin (frequency) found.

- Frequency histogram consists of columns, one for each bin, whose height is determined by the number of samples in the bin.

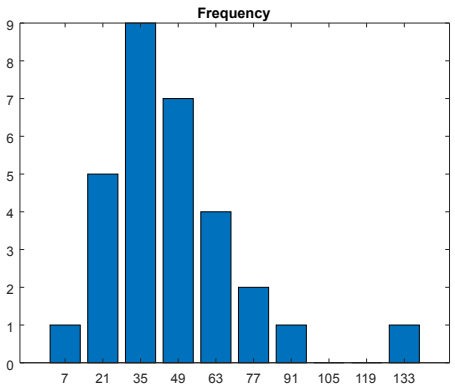
- Relative Frequency histogram is obtained by normalizing Frequency histogram by sample size.

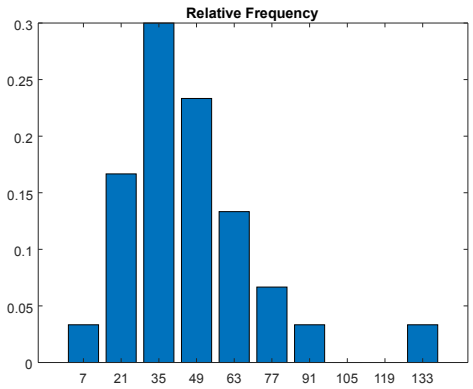
- Shape of histogram depends on
  - number of bins (should not be too few or too many
  - number of bins should increase with sample size
  - number of bins should be chosen to make histogram informative, so that we can

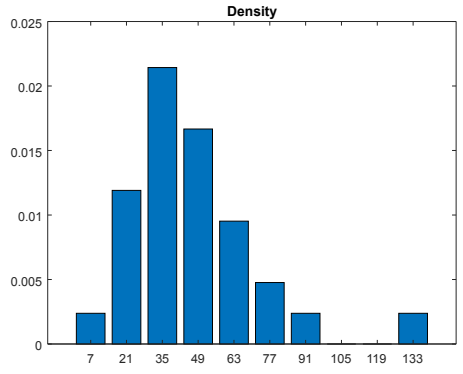
Histogram

Divide samples into $m = 10$ bins:

Interval	Frequency	Relative Frequency	Density
0-14	1	0.033	0.00238
14-28	5	0.167	0.01190
28-42	9	0.30	0.02143
42-56	7	0.233	0.01667
56-70	4	0.133	0.00952
70-84	2	0.067	0.00476
84-98	1	0.033	0.00238
98-112	0	0	0
112-126	0	0	0
126-140	1	0.033	0.00238

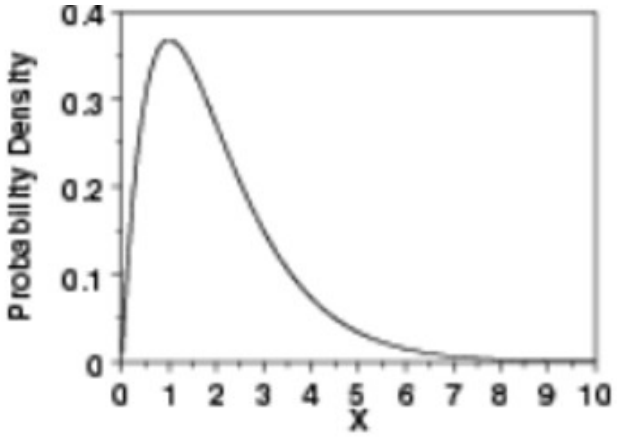


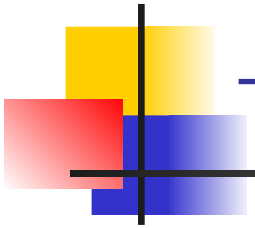




Observations

- Data comes from distribution that is continuous, unimodal, and skew to the right.
- The time 139 stands alone suggesting (it is an outlier).
- There is no indication of heterogeneity; all data except $x = 139$ form rather homogeneous group that fits a gamma distribution.





Typical Questions

- Which values are the most common? Why?
- Which values are rare? Why? Does that match your expectations?
- Can you see any unusual patterns? What might explain them?

- Clusters of similar values suggest that subgroups exist in your data. To understand the subgroups, ask:

  - How are the observations within each cluster similar to each other?

  - How are the observations in separate clusters different from each other?

  - How can you explain or describe the clusters?

  - Why might the appearance of clusters be misleading?

Box Plot

- Sample Mean: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1447}{30} = 48.233$
- Sample Median: $\tilde{X} = \frac{\{X_{(15)} + X_{(16)}\}}{2} = \frac{(42+43)}{2} = 42.5$
- Quartiles divide the ordered sample data into 4 equal (probability)
  Q_1 First Quartile
  Q_2 Second Quartile (Median)
  Q_3 Third Quartile
- Ordered Samples from CPU data:

9	15	19	22	24	25	30	**34**	35	35	36	36	37	38	**42**
43	46	48	54	55	56	56	**59**	62	69	70	82	82	89	139

first quartile

third quartile

- Boxplot graphically depicts the main descriptive statistics of numerical data through their quartiles.

Box Plot

- Boxplot graphically depicts the main descriptive statistics of numerical data through their quartiles.

Five-point summary: $\left(\min X,\ Q_1,\ \widetilde{X},\ Q_3,\ \max X\ \right)$

Interquartile Range (IQR): $Q_3 - Q_1$

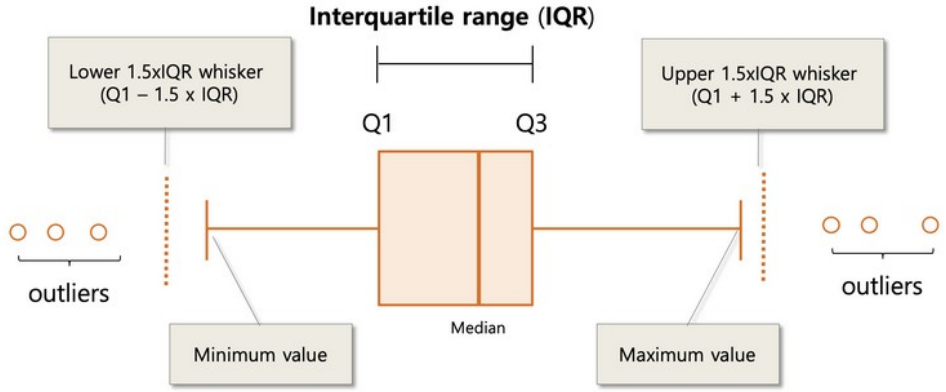
- represents a measure of statistical dispersion
- it represents how 50% of the points are dispersed.
- outliers are defined as samples that below $(Q_1 - 1.5\text{IQR})$ or above $(Q_3 + 1.5\text{IQR})$

- To construct the boxplot

Draw a box between the first and third quartiles: $Q_1(25\%)$ and $Q_3(75\%)$ position

Draw a line inside the box for the median and extend whiskers to the largest sample 1.5IQR above the third quartile, and the smallest sample that is no more that 1.5IQR below the first quartile.

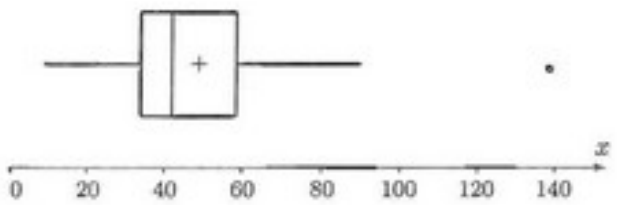
Box Plot



- From the CPU data

Mean: $\bar{X} = 48.233$
Median: $\tilde{X} = 42.5$
25% percentile: $Q_1 = 34$
75% percentile: $Q_3 = 59$
Minimum sample: $\min X = 9$
Maximum sample: $\max X = 139$
$1.5\text{IQR} = 1.5(59 - 34) = 37.5$
$Q_1 - 1.5\text{IQR} = -3.5$
$Q_3 + 1.5\text{IQR} = 96.5$

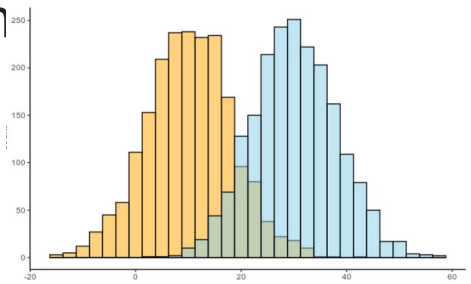
A boxplot of the CPU data:



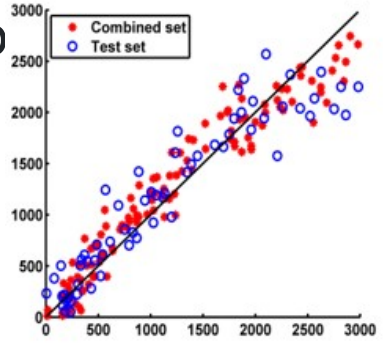
- ✓ Maximum sample not more than $(Q_3 + 1.5\text{IQR})$ is 89. Therefore, 139 represents outlier.
- ✓ Minimum sample is more than $(Q_1 - 1.5\text{IQR})$. No outlier is observed at lower end.

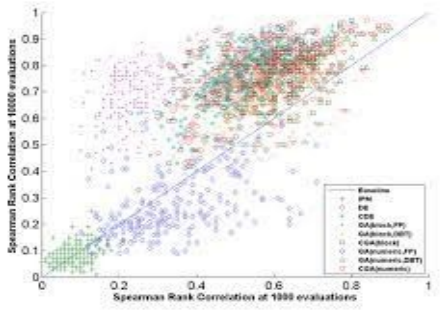
Data Visualization

- Histogram

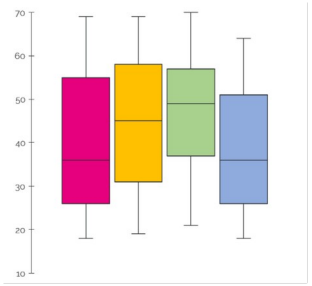


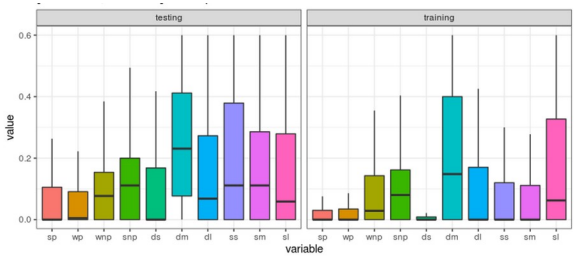
- Scatter Plo





- Box Plot

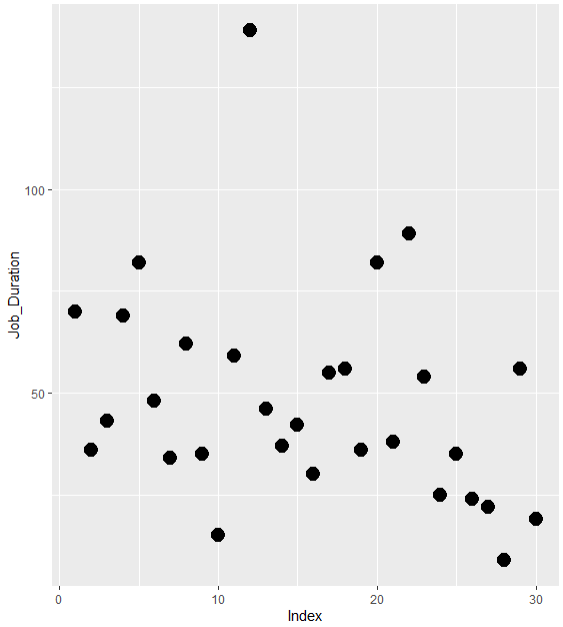


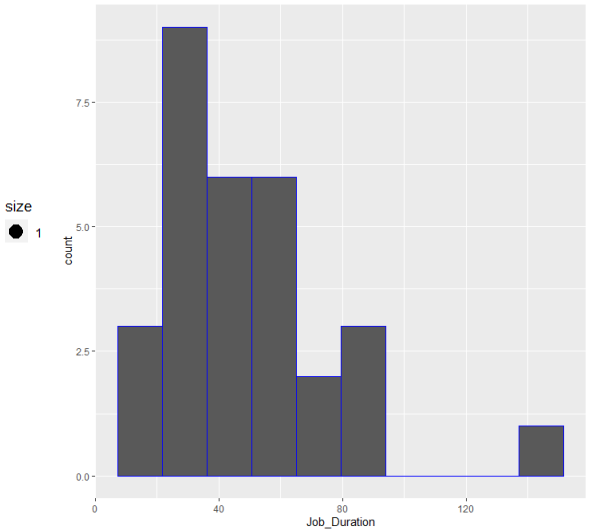


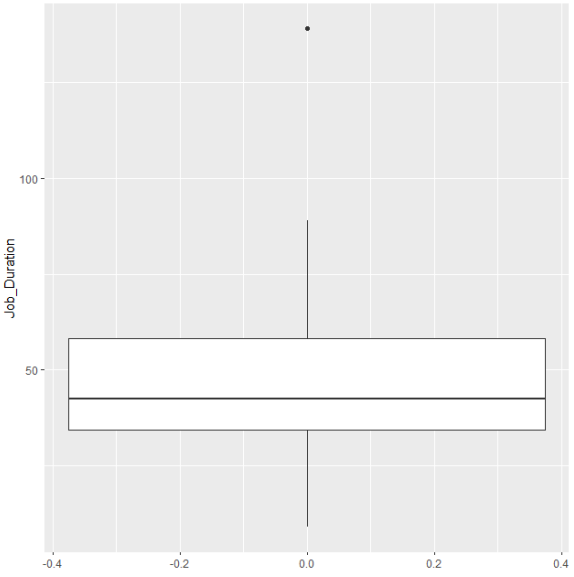
R – Codes for CPU Time example

```
install.packages("tidyverse")
library(tidyverse)
ggplot2::ggplot()
ggplot2::mpg

x1 = c(70, 36, 43, 69, 82, 48, 34, 62, 35, 15, 59, 139, 46, 37, 42,
+      30, 55, 56, 36, 82, 38, 89, 54, 25, 35, 24, 22, 9, 56, 19)

Jobs = data.frame(x1)
colnames(Jobs) <- "Job_Duration"

## Scatter Plot
ggplot(data = Jobs) +
  geom_point(mapping = aes(x=Index, y = Job_Duration, size=1))

## Histogram
ggplot(Jobs, aes(x=Job_Duration)) + geom_histogram(bins=10,color="blue")

## Box Plot
ggplot(data = Jobs, mapping = aes(x=Index, y = Job_Duration, size=1)) +
  geom_boxplot()
```

Data Visualization for CPU Time example







R – Codes for **mpg** dataset

```
install.packages("tidyverse")
library(tidyverse)
ggplot2::ggplot()
ggplot2::mpg

## Scatter Plot
ggplot(data = mpg) +
   geom_point(mapping = aes(x = displ, y = hwy, color=class))

## Histogram
ggplot(data = mpg) +
  geom_bar(mapping = aes(x =class))


ggplot(data = mpg) +
   geom_bar(mapping = aes(x = displ, color=class))

## Box Plot
ggplot(data = mpg, mapping = aes(x = class, y = hwy)) +
   geom_boxplot()

 ggplot(data = mpg, mapping = aes(x = class, y = hwy)) +
   geom_boxplot() +  coord_flip()
```

Data Visualization for **mpg** dataset