Lecture Notes - Data Types & Visualization

Curry W. Hilton

Data Types

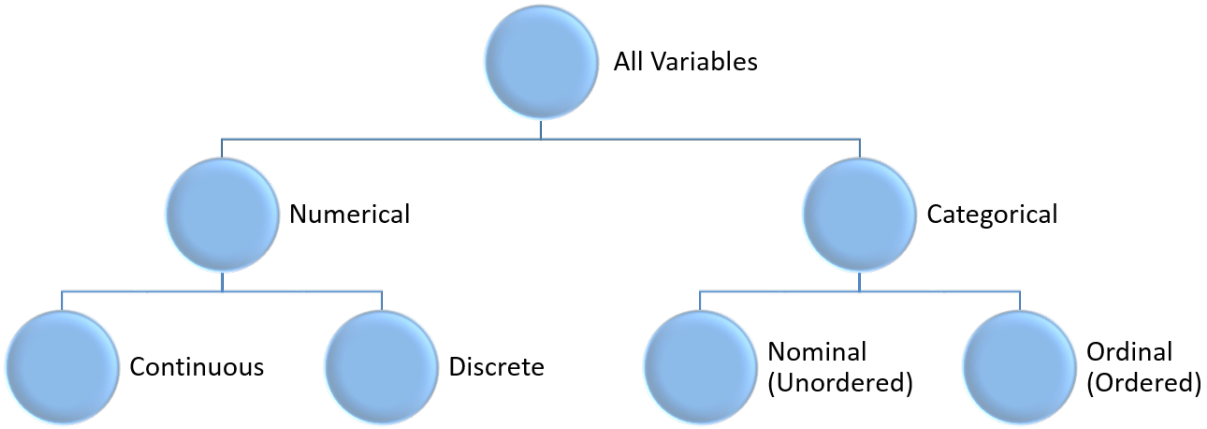

Figure 1:

Numerical Data (Quantitative)

Data that can be measured and recorded in "number" form

- Can perform any type of mathematical operation on numerical data
- Data can be sorted in descending/ascending order
- Data can be expresed in decimal or fractions

Continuous Numerical Data

The possible values of continuous numerical data are considered "uncountable" on a given interval. For example, on the interval [1,5] inclusive, there are an infinite amount of possible data values. Such as: 1, 5, 2, 2.5, 2.7, 3.456, 4.667859402, etc.

- Examples:
  - The bar tabs on any given night a Rounders (could be argued as discrete but close enough to continuous)
  - The amount of gasoline in gallons each customer pumped in a day at the Shell station
  - The number of chips one could consume during an Alabama Football game (assumming fractions of chips could be consumed)

```
library(datasets)   # load "datasets" package
View(ToothGrowth)   # View "ToothGrowth" dataset
```

Discrete Numerical Data

The possible values of discrete numerical data are considered "countable" in a list. For example, in a list from 2 - 5, there are 4 possible values {2, 3, 4, 5}.

- Examples:

- The population of counties in Alabama
- The number of potatoes used each day at the local Five Guys
- The shoe size of students in Bidgood Hall

```
library(datasets)   # load "datasets" package
View(AirPassengers) # View "AirPassengers" dataset
```

Numerical Data Analysis - Univariate

Descriptive Statistics

Mean

Sum of all observations of the numerical variable of interest divided by the number of observations ($\bar{x}$ or μ depending on sample or population).

$$\mu = \frac{1}{N}\sum_{i=1}^{N}(x_1...x_n)$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n}(x_1...x_n)$$

```
mean(AirPassengers)   # mean function called on the atomic (only variable in dataset)
```

```
## [1] 280.2986
```

```
                      # variable in "AirPassengers" dataset
```

Median

When data is ordered in ascending order, the number in the middle is considered the median value.

```
median(AirPassengers) # median function called
```

```
## [1] 265.5
```

Max & Min

Maximum and Minimum value in a set of observations... pretty self explanatory

```
max(AirPassengers)    # max function called
```

```
## [1] 622
```

```
min(AirPassengers)    # min function called
```

```
## [1] 104
```

Variance

Represents the average squared distance from the mean (s^2 => sample variance, σ^2 => population variance).

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

```
var(AirPassengers)   # var function called
```

```
## [1] 14391.92
```

Standard Deviation

Represents the square root of the variance... a means of determining how close the data is from the mean value.

$$\sigma = \sqrt{\sigma^2}$$
$$s = \sqrt{s^2}$$

```
sd(AirPassengers)    # sd function called
```

```
## [1] 119.9663
```

6-Number Summary

In R, the function "summary()" provides a 6-number summary of descriptive statistics, including the Mean, Median, 25th Percentile (Q_1), 75th Percentile (Q_3), Max, and Min.

```
summary(AirPassengers)   # summary function called
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   104.0   180.0   265.5   280.3   360.5   622.0
```

Interquartile Range (IQR)

The difference between the 75th Percentile and 25th Percentile. Another measure of variability in data. Also represents the "length" of the box from the box and whiskers plot (More to come in visualization)

$$IQR = Q_3 - Q_1$$

Numerical Data Visualization - Univariate

Dot Plots

A one-variable scatter plot. The stacked version adds more value. One can infer "balancing" point of distribution of observations... aka the mean value... think about a fulcrum. Also the stacked dot plot displays spread of data and potential outliers.

```
stripchart(beaver1$temp, method="stack", offset=0.5, pch=1)  # dot plot of temperature
                                                             # in "beaver1" dataset
```

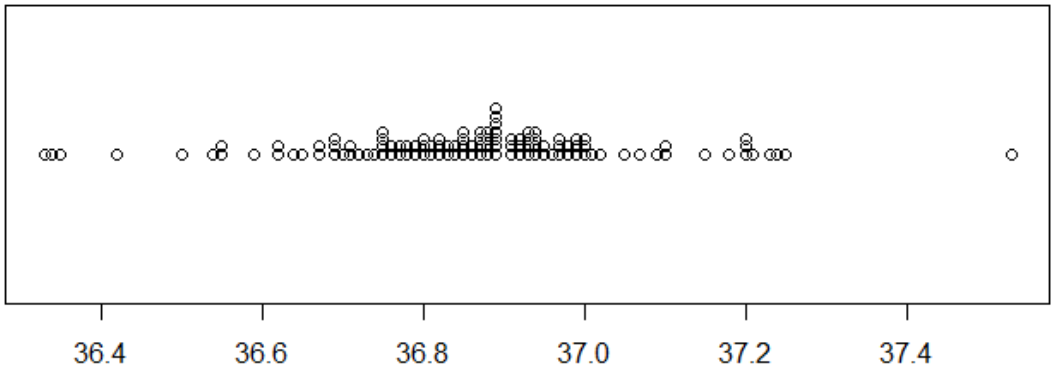

Figure 2:

Histograms

A graphical display of data measuring frequency of observations in predefined bins (intervals of data). Useful in observing the "distribution" or shape of the data and the frequency of data points (Also known as data density).

```
hist(beaver1$temp)    # histogram of temperature in "beaver1" dataset
```

- *Modal Type*: The most frequent observation in a distribution of data... when determining the mode find the most "prominent" peak in the histogram. Modal type can be unimodal, bimodal, and multimodal.
- *Skewness*: Direction of "tail" of data. Measurement interpretation in ()
  - Left Skewed: Long left tail (Negative Skewness)
  - Right Skewed: Long right tail (Positive Skewness)
  - No Skewness: Symmetric tails (Zero Skewness)

```
install.packages("moments")   # install the "moments" package
```

```
library(moments)              # load the moments package
```

```
## Warning: package 'moments' was built under R version 3.2.3
```

```
skewness(beaver1$temp)        # determine the skewness of temperature
```

```
## [1] -0.02782567
```

Box and Whisker Plot

Uses 5 of the 6-number summary statistics, Max, Q_3, Median, Q_1, and Min. It also uses the IQR to compute the "reach" values for the whiskers.

Whisker Calculations: - Upper: $Q_3 + (1.5 * IQR)$ - Lower: $Q_1 - (1.5 * IQR)$

Outlier Detection: Observations that appears extreme relative to the other observations in the dataset.

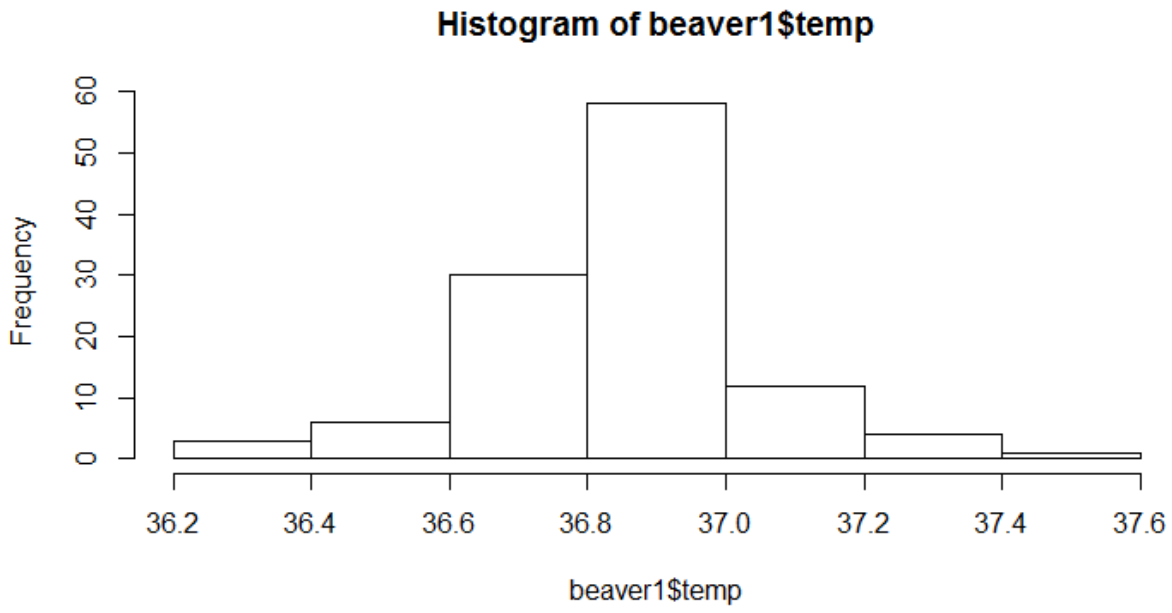

Figure 3:

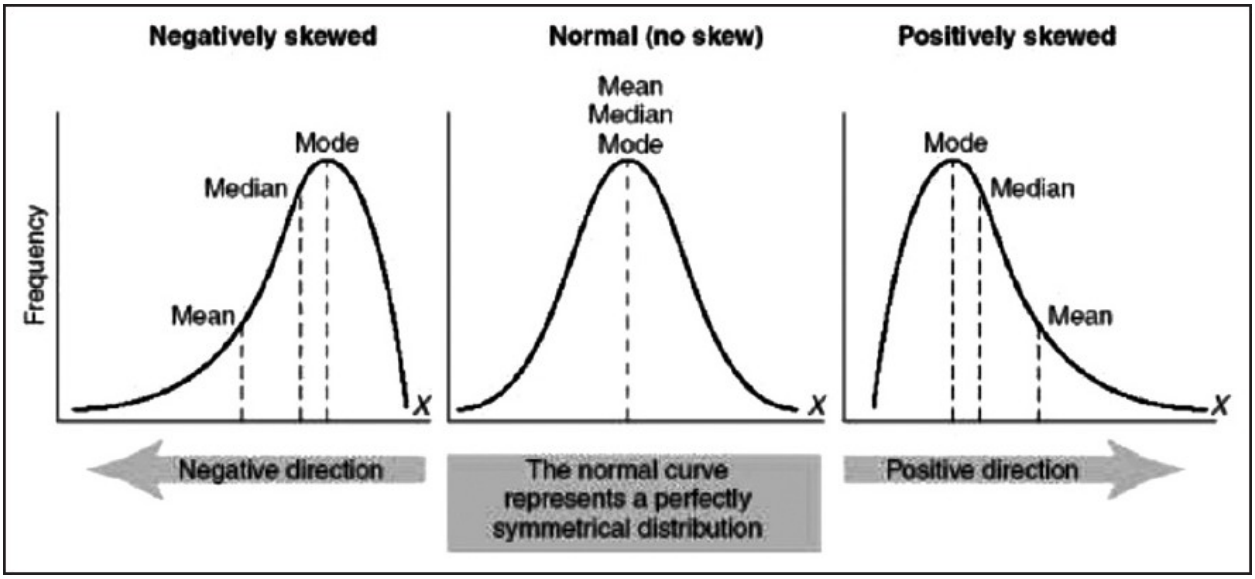

Figure 4:

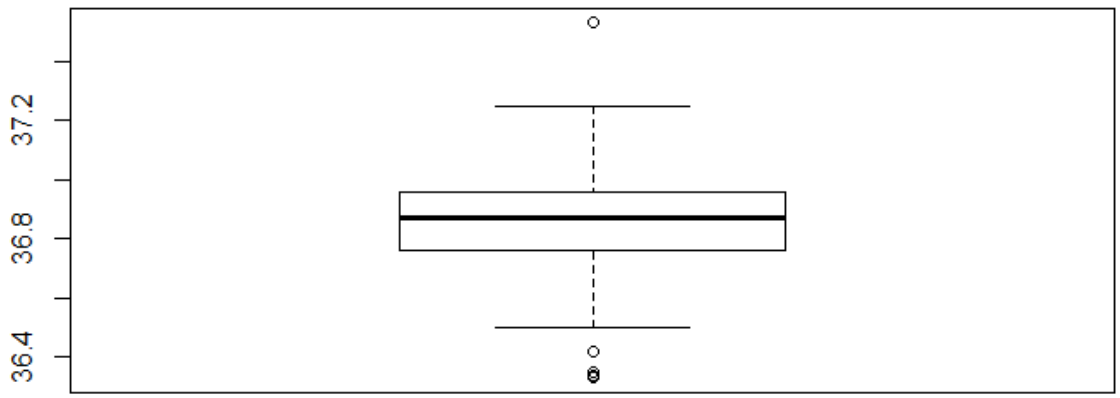

Figure 5:

Numerical Data Analysis - Multivariate

Descriptive Statistics - Measures of Association

Covariance

A descriptive measure of linear association between variables (In our case two variables). Population Covariance => σ_{xy}; Sample Covariance => s_{xy}

$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)$$

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

```
cov(births$weeks, births$weight)  # calculate covariance of two variables
```

Interpretation is tricky using only covariance. One can infer negative or positive relationship based on the sign of the covariance statistic.

Correlation Coefficient

A descriptive measure of linear association between variables that is bounded by -1 and 1. Where the closer to |1| the statistic is the stronger the linear relationship. More frequently used than covariance for interpretation sake. Population Correlation => ρ_{xy}; Sample Correltion => r_{xy}

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

```
cor(births$weeks, births$weight)   # calculate the correlation between two variables
```

Numerical Data Visualization - Multivariate

Scatter Plot

Graphical means to represent the relationship between variables.

```
plot(births$weeks, births$weight)   # create a scatter plot of two variables
```

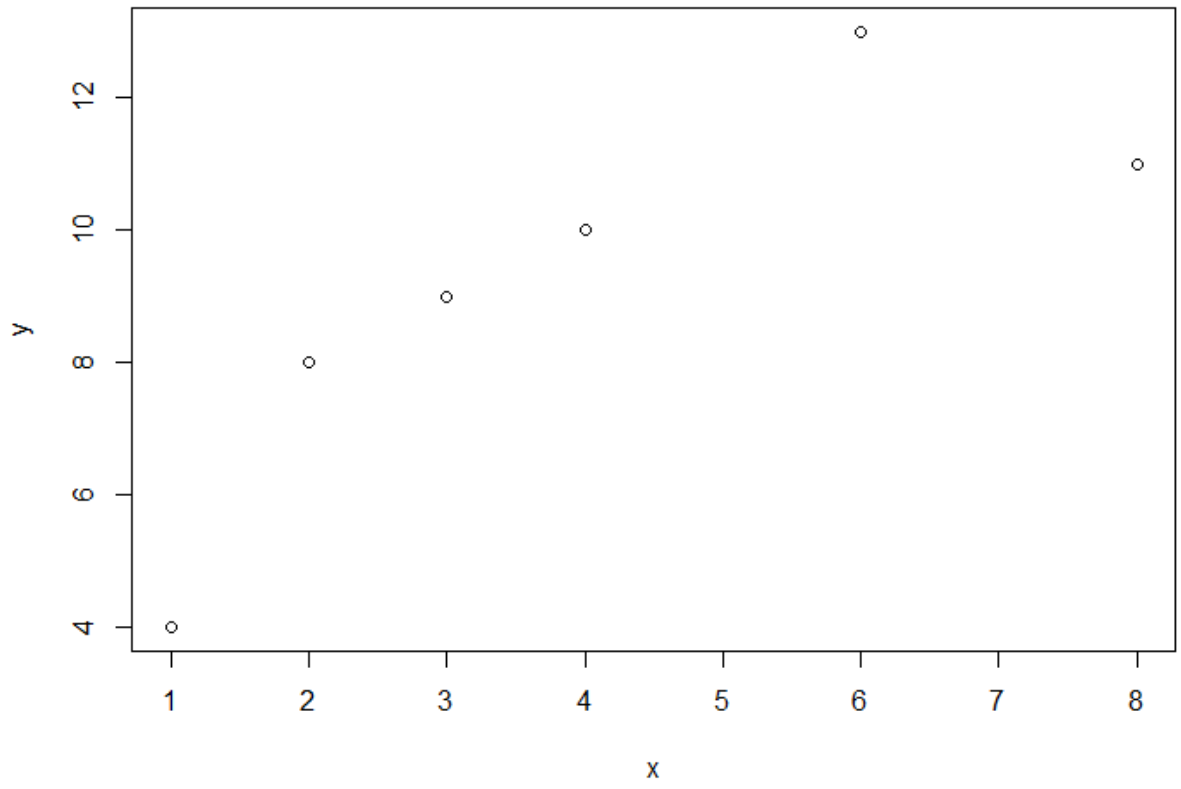

Figure 6:

Categorical Data (Qualitative)

Data that can be grouped by certain attributes.

Nominal Categorical Data (Unordered)

Data that allows assignment of categories but no clear ordering or ranking of groups exist.

- Examples:
  - Gender (Female or Male)
  - Marital Status (Married, Divorced, Single)

```
library(openintro)   # load "openintro" package
View(births)         # View "births" dataset
```

Ordinal Categorical Data (Ordered)

Data that allows assignment of categories with clear ordering or ranking among groups.

- Examples:
  - Education Level (High School, Undergraduate, Graduate)
  - Economics Status (Low, Medium, High)
  - Olympic medals (Gold, Silver, Bronze)

```
library(openintro)    # load "openintro" package
View(mammals)         # View "mammals" dataset
```

Categorical Data Analysis

Frequency Tables & Relative Frequency Tables

Tables that summarize 1 categorical variable in terms of frequency, percentages, or proportions

```
table(iris$Species)   # building a frequency table with count
```

```
##
##     setosa versicolor  virginica
##         50         50         50
```

Contingency Tables

Tables that summarize 2 categorical variables. Usually expressed as frequencies or proportions.

```
library(openintro)
```

```
## Warning: package 'openintro' was built under R version 3.2.3
```

```
## Please visit openintro.org for free statistics materials
```

```
##
## Attaching package: 'openintro'
```

```
## The following object is masked from 'package:datasets':
##
##     cars
```

```
table(births$premature, births$smoke)   # building a contingency table
```

```
##
##             nonsmoker smoker
##   full term        87     42
##   premie           13      8
```

Bar Plots

```
barplot(table(births$smoke))
```

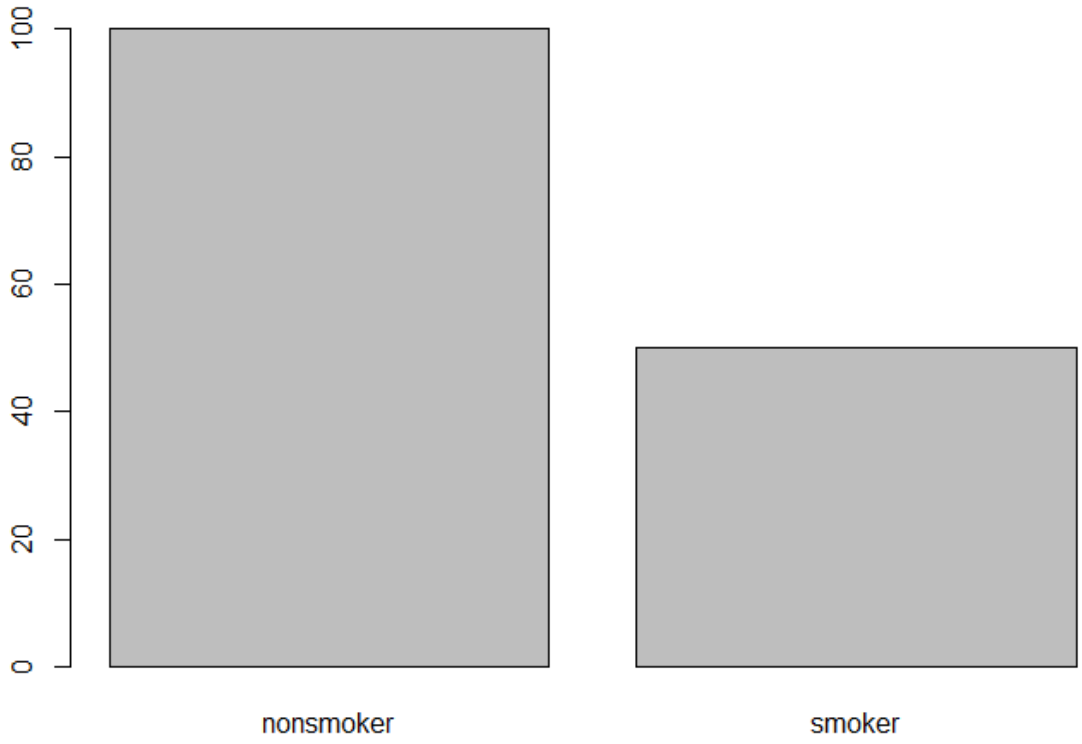

Figure 7:

Mixed Data Type Visualizations

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.2.5
```

```
library(gcookbook)
```

```
## Warning: package 'gcookbook' was built under R version 3.2.5
```

```
##
## Attaching package: 'gcookbook'
```

```
## The following object is masked from 'package:openintro':
##
##     marathon
```

Scatterplot: Two Numerical Variables - Grouped by Categorical Variable

```
qplot(carat, price, data=diamonds,
      colour=color)
```

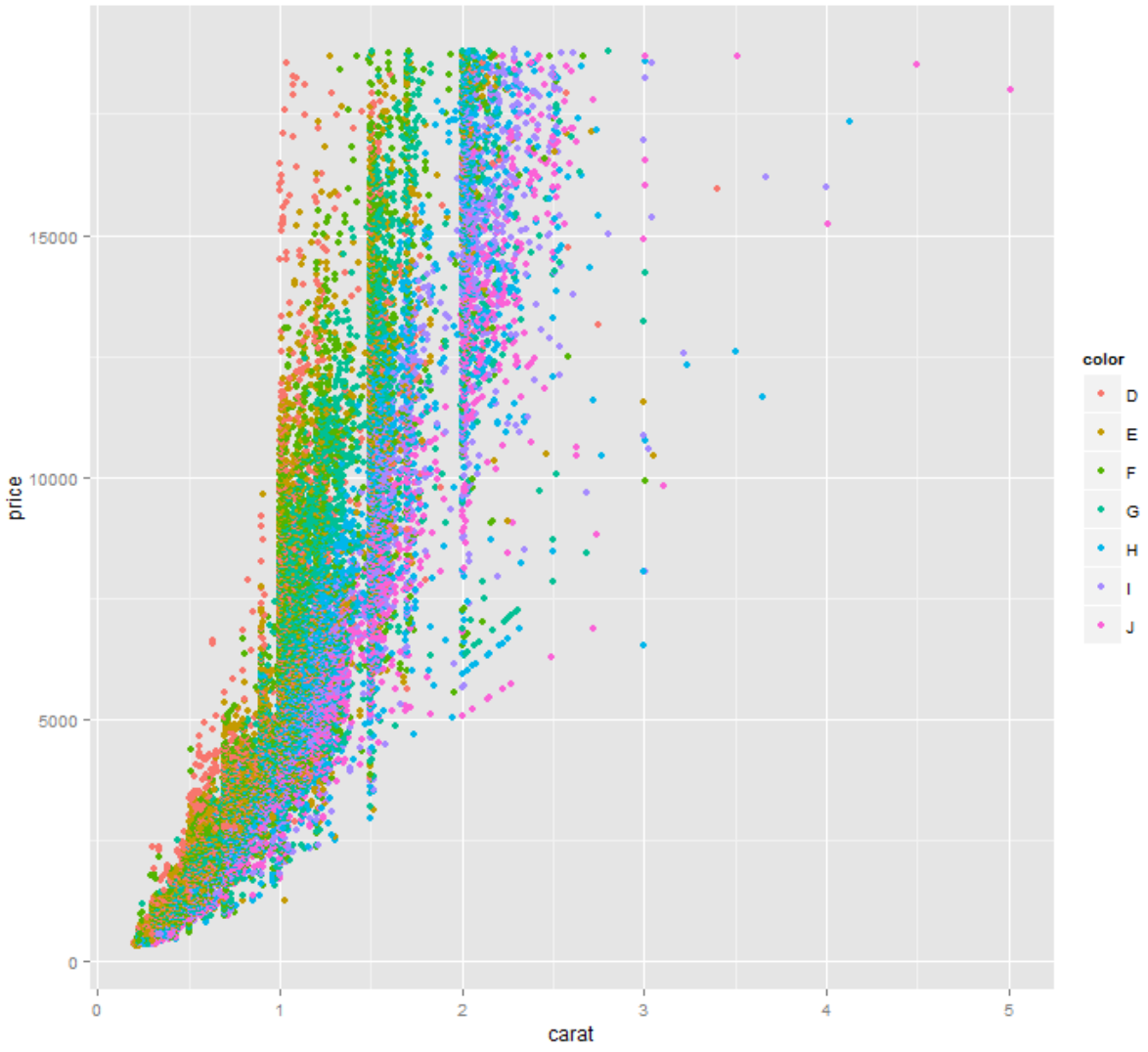

Figure 8:

Box and Whiskers: One Numerical Variable - Grouped by Categorical Variable

```
qplot(factor(cut), price, data = diamonds, geom = "boxplot")
```

Box Plot: One Numerical Variable - Grouped by Two Categorical Variables

```
ggplot(cabbage_exp, aes(x=Date, y=Weight, fill=Cultivar))+geom_bar(position="dodge", colour="black", sta
```

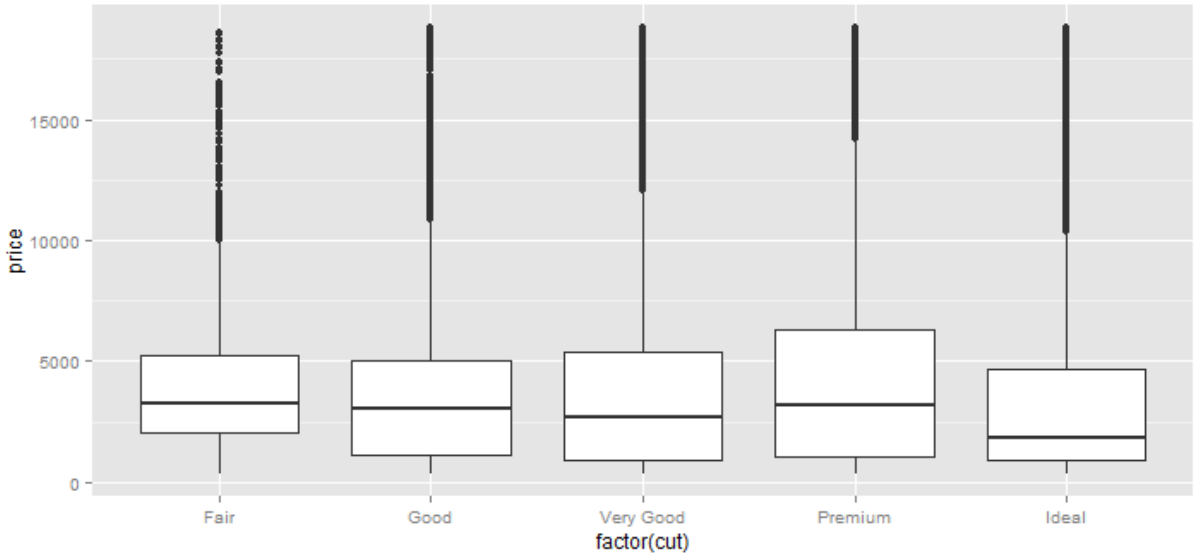

Figure 9:

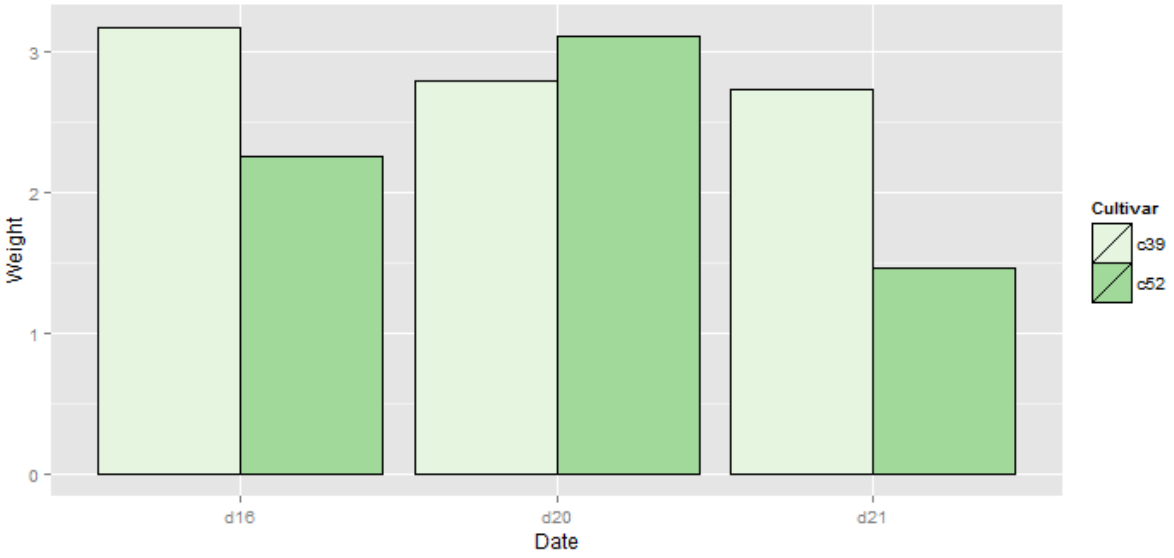

Figure 10:

Correlation Matrix - Multivariate

```
pairs(ChickWeight)
```

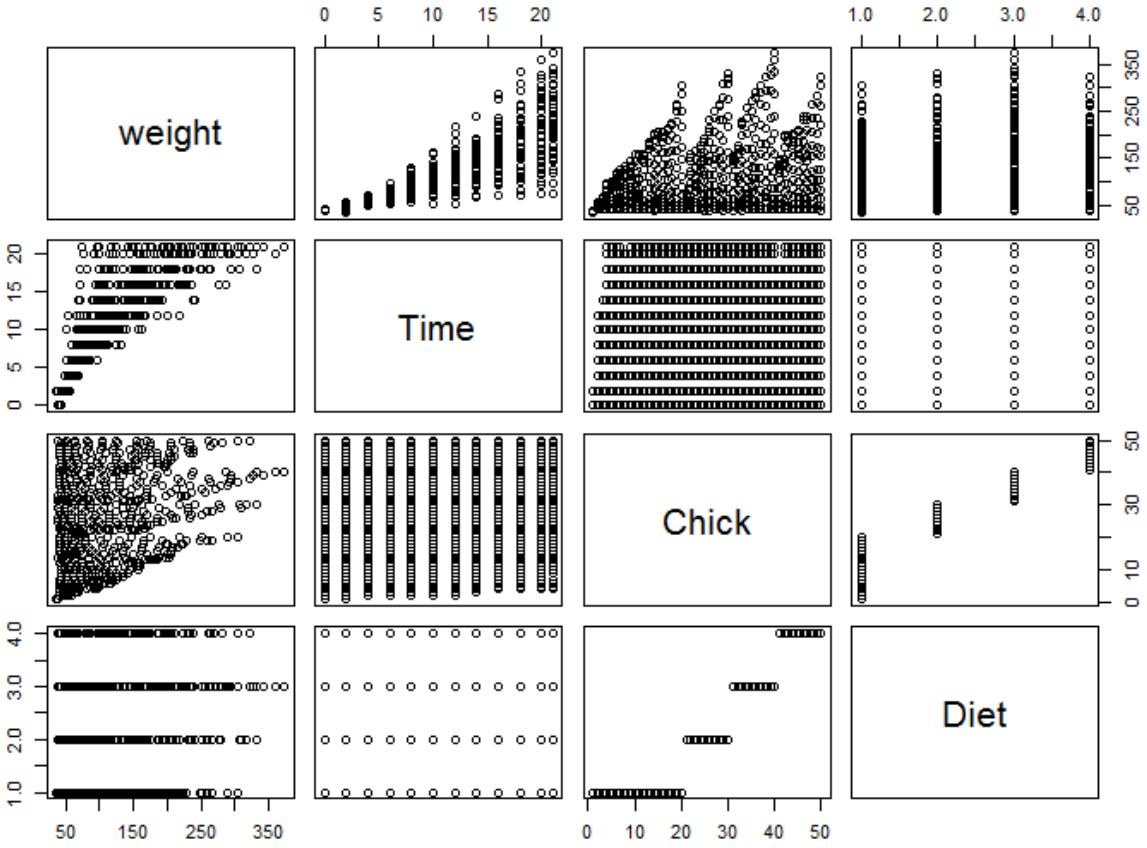

Figure 11:

Distribution Analysis - One Numerical Variable - Grouped by Two Categorical Variables

```
qplot(price, data=diamonds, geom="density", fill=cut, alpha=I(.5),
      main="Distribution of Price", xlab="Price",
      ylab="Density")
```

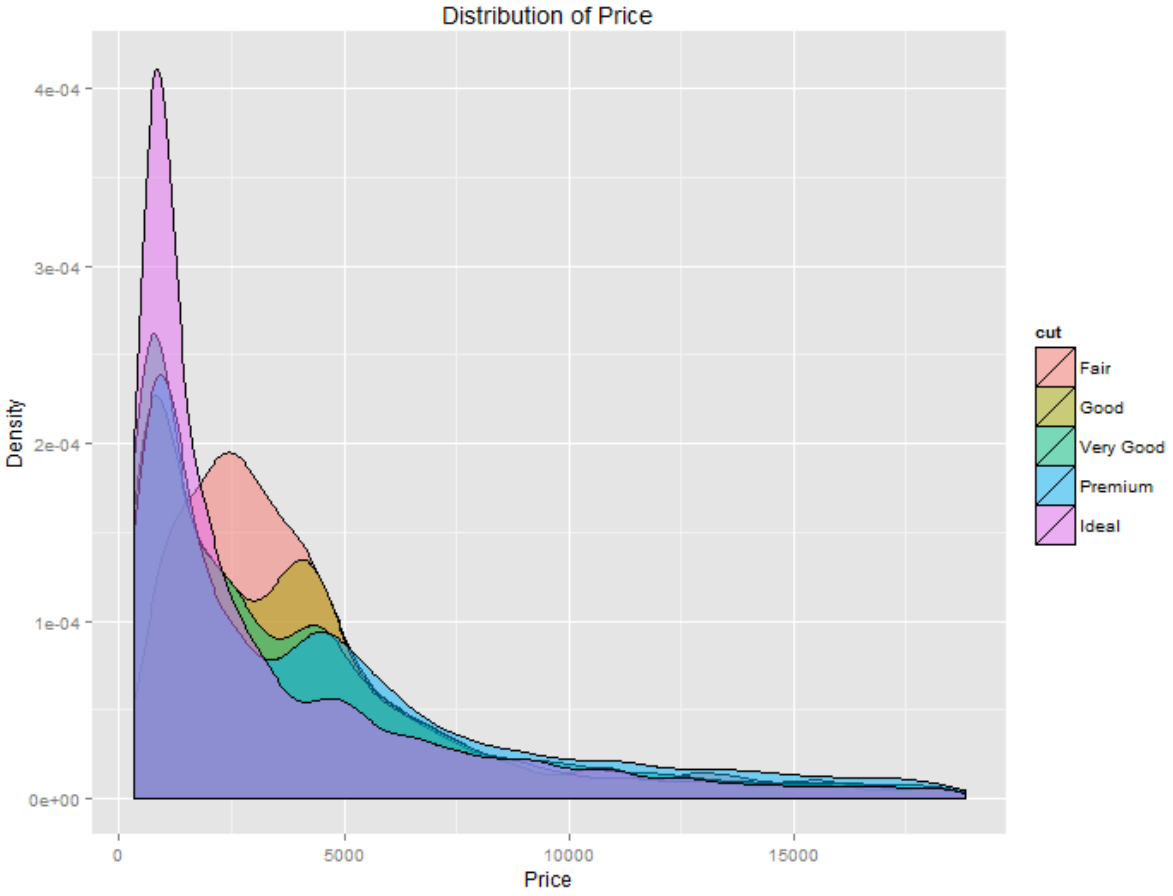
Distribution of Price
Density
4e-04
3e-04
2e-04
1e-04
0e+00
0
5000
10000
15000
Price
cut
Fair
Good
Very Good
Premium
Ideal

Figure 12: