Working with Data in R

Instructor: Mary Yang, Ph.D.

UALR/UAMS Joint Bioinformatics Program
University of Arkansas Little Rock

June 17, 2021

Outline

- Quick review of data type and data structure in R
- Introduce to ggplot
- Build prediction models in R
- Introduce to R Markdown

What is R?

- R is object-oriented, open source programming language
- R is an integrated suite of software facilities for data manipulation, simulation, calculation and graphical display
- R runs on a wide variety of platforms such as Linux, Windows, and macOS
- The R project web page
  - **http://www.r-project.org**
- RStudio is a convenient interface and allows the user to run R in a more user-friendly environment
  - **http://www.rstudio.com/products/rstudio/download/**

R Data Types

Types	Examples
Integer: Natural numbers	1, 2, 3
Numeric: Decimal values	1.5, 2.2, 3.7
Logical: Boolean values	TRUE or FALSE (T or F)
Character: Text or string values	"a" "cat" "blue"

Data Structures

- R has a wide variety of data structures:
  - vector
  - matrix
  - data frame
  - list

	Homogeneous	Heterogeneous
1d	Vector	List
2d	Matrix	Data frame
nd	Array	

Factor

- R has a special data structure for categorical data, called factor.
- Factors are important for statistical analysis and for plotting.
- Internally a factor is stored as a numeric value associated with each level.

```
> Gender <- factor(c("male", "female", "female", "male"))
> Gender
[1] male female female male
Levels: female male

> mode(Gender)
[1] "numeric"

> str(Gender)
 Factor w/ 2 levels "female","male": 2 1 1 2
```

Advanced graphics: ggplot2

- ggplot2 is a plotting system based on the Grammar of Graphics and expands the capabilities of base R graphics system.
- ggplot2 is designed to work in a layered fashion, starting with a layer showing the raw data then adding layers of annotation and statistical summaries.
- To add a layer, use + operator.
- We only need minimal changes if the underlying data change or if we decide to change from a bar plot to a scatterplot.

```
# Install ggplot2 package
> install.packages("ggplot2")

# load the ggplot2 package
> library(ggplot2)
```

ggplot2

- The following basic template that can be used for different types of plots:

```
ggplot(data = <DATA>, mapping = aes(<MAPPINGS>)) + <GEOM_
    FUNCTION>()
```

- Every graphic made by ggplot2 have at least one aesthetic (aes()) and at least one geom (layer).
  - **aes():**The aesthetic maps your data to your geometry (layer).
  - **geometry layer)** geometry specifies the type of plot you are making (point, line, bar, etc.)
- ggplot2 offers many different geom_functions. The most common one including
  - **geom_point()** for scatter plots, dot plots, etc.
  - **geom_boxplot()** for boxplot
  - **geom_line()** for trend lines, time series, etc
  - **geom_bar()** for bar plots, dot plots, etc.
  - **geom_histogram** fo histogram plot

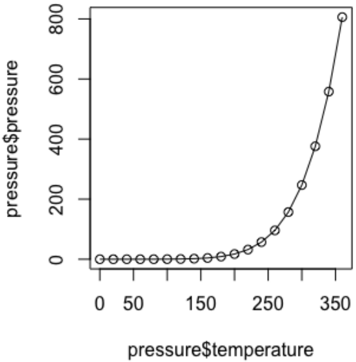
Line graph

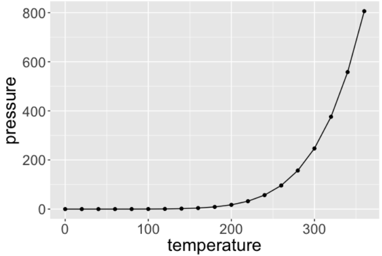
```
> ?pressure
> str(pressure)
'data.frame': 19 obs. of 2 variables:
 $ temperature: num 0 20 40 60 80 100 120 140 160 180 ...
 $ pressure   : num 0.0002 0.0012 0.006 0.03 0.09 0.27 0.75 1.85
    4.2 8.8 ...
```

pressure: data on the relation between temperature in degrees Celsius and vapor pressure of mercury in millimeters (of mercury).

Question: How to show the presssure change versus temperature?

Basic plot vs ggplot: line graph





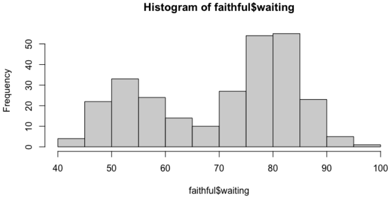

```
# Generate the left figure
> plot(pressure$temperature, pressure$pressure, type="l")
> points(pressure$temperature, pressure$pressure)

#Generate the right figure
ggplot(pressure, aes(x=temperature, y=pressure)) + geom_line() + geom_point() + theme(
     text = element_text(size=20))
```

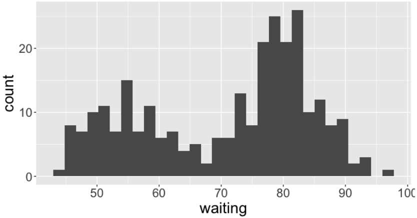
```
> ?faithful
> str(faithful)
'data.frame': 272 obs. of 2 variables:
 $ eruptions: num 3.6 1.8 3.33 2.28 4.53 ...
 $ waiting : num 79 54 74 62 85 55 88 85 51 85 ...
```

```
Description
Waiting time between eruptions and the duration
of the eruption for the Old Faithful geyser
in Yellowstone National Park, Wyoming, USA.
```

Question: Plot distribution of waiting time

Basic plot vs ggplot: histogram






```
# Generate the left figure
hist(faithful$waiting)

#Generate the right figure
ggplot(faithful, aes(x=waiting)) + geom_histogram() + theme(text = element_text(size=20))
```

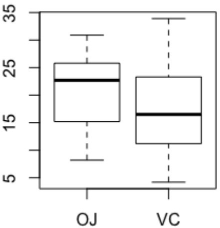
Boxplot

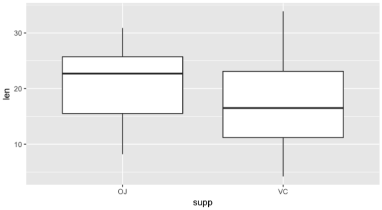
```
> ?ToothGrowth
> str(ToothGrowth)
'data.frame': 60 obs. of 3 variables:
 $ len : num 4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
 $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
 $ dose: num 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

```
The response is the length of odontoblasts (cells responsible for tooth
in 60 guinea pigs. Each animal received one of three dose levels
of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods,
orange juice or ascorbic acid (a form of vitamin C and coded as VC).
```

Question: How to compare tooth growth of pigs under different conditions?

Boxplot

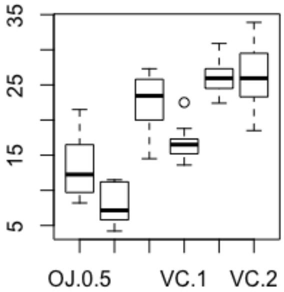


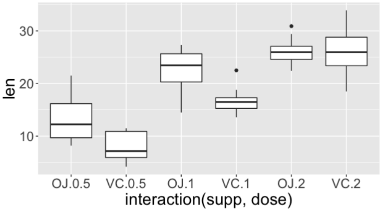



```
# Left figure
> plot(ToothGrowth$supp, ToothGrowth$len)
# Formula syntax
> boxplot(len ~ supp, data = ToothGrowth)

# Right figure
> ggplot(ToothGrowth, aes(x=supp, y=len)) + geom_boxplot()
```

Question: Is this tooth growth pattern for the two dlievery methods same for all the three dose levels?

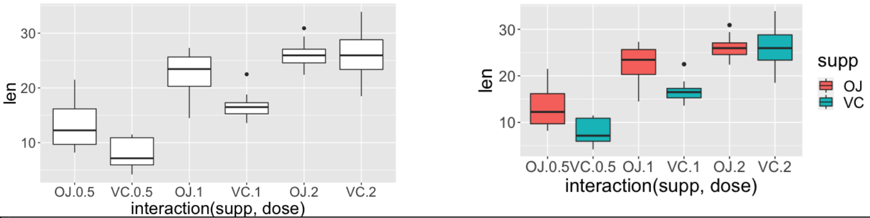
Boxplot






```
# Left figure
#Put interaction of two variables on x−axis
>boxplot(len ˜ supp + dose, data = ToothGrowth)

# Right figure
> ggplot(ToothGrowth, aes(x=interaction(supp, dose), y=len)) + geom_
    boxplot() + theme(text = element_text(size=20))
```

Boxplot



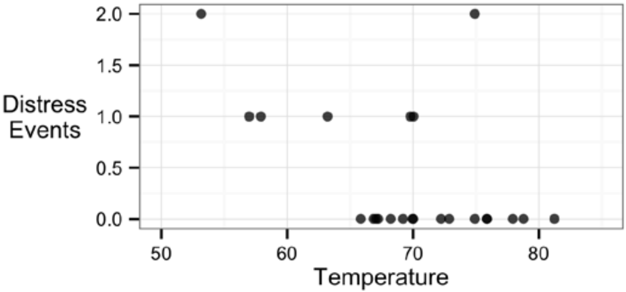

```
# Left figure
> ggplot(ToothGrowth, aes(x=interaction(supp, dose), y=len)) +
      geom_boxplot() +
     theme(text = element_text(size=20))

# Right figure
> ggplot(ToothGrowth, aes(x=interaction(supp, dose), y=len, fill=supp)) +
      geom_boxplot() +
      theme(text = element_text(size=20))
```

Build prediction models using R

- Preding sale using marketing data
- Predicting medical cost using health insurance data

Data

- On January 28, 1986, Space Shuttle Challenger broke apart when a rocket booster failed caused by the failure of O-ring seals.
- 23 shuttle launches which recorded the number of O-ring failures versus the launch temperature



Question: Can we predit O-ring failure?

Data trend

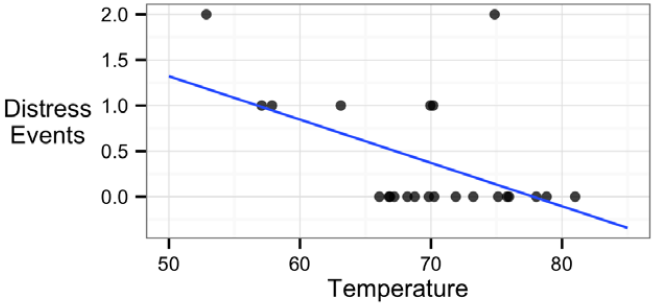
- There is an apparent trend between temperature and number of failures.
- Launches occurring at higher temperatures tend to have fewer O-ring failures.
- Linear regression model defines the relationship between a **dependent variable(x)** and a single **independent predictor variable (y)**, **using a line** denoted by an equation in the following form:

$$y = \alpha + \beta x \tag{1}$$

Linear regression model

Suppose we obtained this equation to fit the data

$$y = 4.30 - 0.057x \quad (2)$$



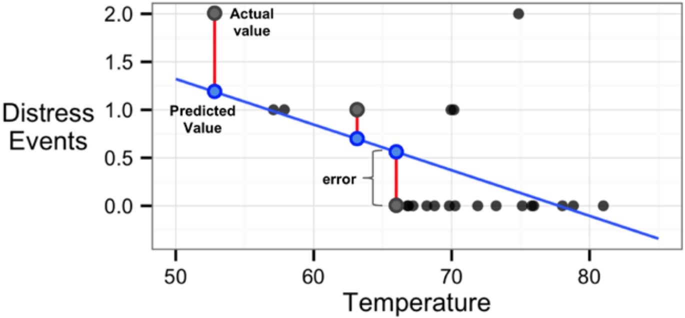
How to interpret the model

Suppose we obtained this equation to fit the data

$$y = 4.30 - 0.057x \tag{3}$$

- At 60 degrees Fahrenheit, we predict just under one O-ring failure.
- At 70 degrees Fahrenheit, we expect around 0.3 failures.
- If we extrapolate our model all the way out to 31 degrees −the forecasted temperature for the Challenger launch, we would expect about $4.30 - 0.057 * 31 = 2.53$ O-ring failures

Ordinary least squares estimation



$$min \sum (y_i - \hat{y})^2 = \sum e_i^2 \quad (4)$$

Marketing data

```
> install.packages("datarium")
> library(datarium)
> ?marketing
> str(marketing)
'data.frame': 200 obs. of 4 variables:
 $ youtube : num 276.1 53.4 20.6 181.8 217 ...
 $ facebook : num 45.4 47.2 55.1 49.6 13 ...
 $ newspaper: num 83 54.1 83.2 70.2 70.1 ...
 $ sales : num 26.5 12.5 11.2 22.2 15.5 ...
```

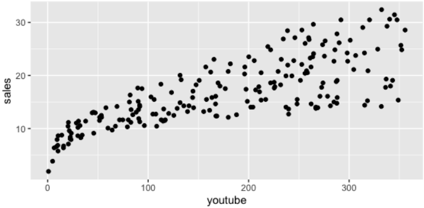
```
A data frame containing the impact of three advertising medias
(youtube, facebook and newspaper)
 on sales. Data are the advertising budget in thousands of dollars
 along with the sales
```

Question: Can we predict sale using advertising budget?

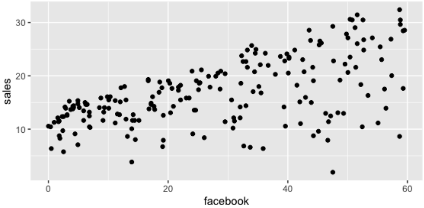
Visualization

- Create a scatter plot displaying the sales units versus youtube and facebook advertising budget

```
> ggplot(marketing, aes(x = youtube, y = sales)) + geom_point()
```




```
> ggplot(marketing, aes(x = facebook, y = sales)) + geom_point()
```



Regression model syntax

using the `lm()` function in the `stats` package

Building the model:

```
m <- lm(dv ~ iv, data = mydata)
```

- `dv` is the dependent variable in the `mydata` data frame to be modeled
- `iv` is an R formula specifying the independent variables in the `mydata` data frame to use in the model
- `data` specifies the data frame in which the `dv` and `iv` variables can be found

The function will return a regression model object that can be used to make predictions. Interactions between independent variables can be specified using the * operator.

Making predictions:

```
p <- predict(m, test)
```

- `m` is a model trained by the `lm()` function
- `test` is a data frame containing test data with the same features as the training data used to build the model.

The function will return a vector of predicted values.

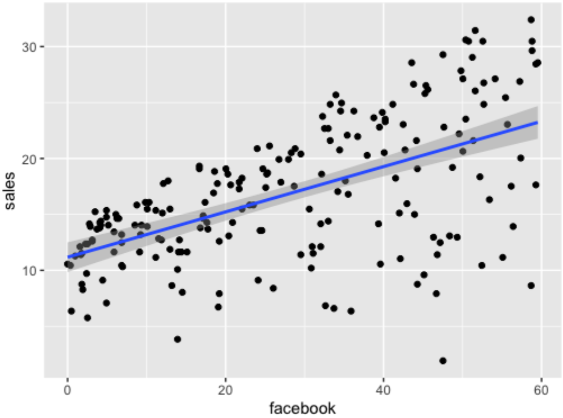
A simple linear regression model

```
> model <- lm(sales ~ youtube, data = marketing)
> model

Call:
lm(formula = sales ~ youtube, data = marketing)

Coefficients:
(Intercept) youtube
    8.43911 0.04754
```

Plot linear regression line

$y = 8.439 - 0.0475x$

```
> ggplot(marketing, aes(, sales)) + geom_point() + stat_smooth(method = lm)
```



View the model

```
> summary(model)

Call:
lm(formula = sales ~ youtube, data = marketing)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.0632  -2.3454  -0.2295   2.4805   8.6548 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 8.439112   0.549412   15.36   <2e-16 ***
youtube     0.047537   0.002691   17.67   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.91 on 198 degrees of freedom
Multiple R-squared:  0.6119,Adjusted R-squared:  0.6099 
F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

View the model

The summary shows six components:

- **Call**: Shows the function call used to compute the regression model.
- **Residuals**: The Residuals section provides summary statistics for the errors in our predictions
- **Coefficients**: Shows the regression beta coefficients and their statistical significance.
  - Predictor variables, that are significantly associated to the outcome variable, are marked by stars.
- **Residual standard error** (RSE), **R-squared** (R2) and the **F-statistic** are metrics that are used to check how well the model fits to our data.

Model accuracy

- The R-squared (R2) ranges from 0 to 1 and represents the proportion of information in the data that can be explained by the model. The adjusted R-squared adjusts for the degrees of freedom.
    - Multiple R-squared: 0.6119, Adjusted R-squared: 0.6099
- The F-statistic gives the overall significance of the model.
    - F-statistic: 312.1 on 1 and 198 DF, $p - value :< 2.2e - 16$

Exercise 1:

- Bulid a linear regression model tp predict the sale using facebook advertisement budget
- Plot a linear regression line to fit the sale data
- Is this a good model?

Multiple linear regression

Question: Can we imporve prediction?

- Most real-world analyses have more than one independent variable.
- Multiple regression equations generally follow the form of the following equation.

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_i x_i + \epsilon \tag{5}$$

- An error term ϵ has been added here as a reminder that the predictions are not perfect.

A multiple linear regression model

```
> model <- lm(sales ~ youtube + facebook + newspaper, data = marketing)
> model

Call:
lm(formula = sales ~ youtube + facebook + newspaper, data = marketing)

Coefficients:
(Intercept) youtube facebook newspaper
   3.526667 0.045765 0.188530 -0.001037
```

View the model

```
> summary(model)

Call:
lm(formula = sales ~ youtube + facebook + newspaper, data = marketing)

Residuals:
     Min       1Q   Median       3Q      Max
-10.5932  -1.0690   0.2902   1.4272   3.3951

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.526667   0.374290   9.422   <2e-16 ***
youtube      0.045765   0.001395  32.809   <2e-16 ***
facebook     0.188530   0.008611  21.893   <2e-16 ***
newspaper   -0.001037   0.005871  -0.177     0.86
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.023 on 196 degrees of freedom
Multiple R-squared:  0.8972,Adjusted R-squared:  0.8956
```

Simualted Medical Insurance Data

- For this analysis, we will use a simulated dataset containing medical expenses for patients in the United States.
- It was created using demographic statistics from the U.S. Census Bureau, and thus approximately reflect real-world conditions.
- The goal of the following analysis is to **use patient data to estimate the average medical care expenses for such population segments.**

Simualted Medical Insurance Data

- The *insurance.csv* file includes 1,338 examples of beneficiaries currently enrolled in the insurance plan
- The features indicating characteristics of the patient as well as the total medical expenses charged to the plan for the calendar year.

Import data

```
> insurance <- read.csv("insurance.csv")
> str(insurance)
'data.frame': 1338 obs. of 7 variables:
 $ age : int 19 18 28 33 32 31 46 37 37 60 ...
 $ gender : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2
     1 ...
 $ bmi : num 27.9 33.8 33 22.7 28.9 ...
 $ children: int 0 1 3 0 0 0 1 3 2 0 ...
 $ smoker : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1
    ...
 $ region : Factor w/ 4 levels "northeast","northwest",..: 4 3 3
    2 2 3 3 2 1 2 ...
 $ charges : num 16885 1726 4449 21984 3867 ...
```

Exlpore data

- It is important to give some thought to how these variables may be related to billed medical expenses.
- For instance, we might expect that older people and smokers are at higher risk of large medical expenses.
- In regression analysis, the relationships among the features are typically specified by the user rather than detected automatically.

Charge distribution

Let's take a look to see how *charge* is distributed

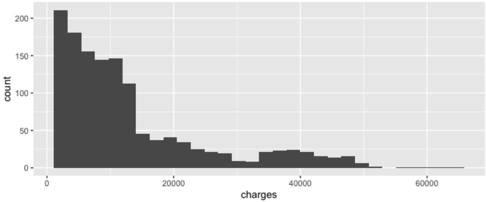
```
> summary(insurance$charges)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1122    4740    9382   13270   16640   63770
```

Here, mean is larger than median.
Question: What kind of distribution of insurance charge you expect?

Dependent variable: charges

Because the mean value is greater than the median, this implies that the distribution of insurance **charges** is right-skewed

```
> ggplot(insurance, aes(charges)) + geom_histogram()
```



Exerise 2: Draw a boxplot of medical charges in different regions.

Dependent categorical variable

- The **gender** variable is divided into male and female levels
- **smoker** is divided into yes and no.
- From the **str(insurance)** we know that region has four levels,
- Let's take a closer look to see how they are distributed.

```
> table(insurance$region)
northeast northwest southeast southwest
      324       325       364       325
```

Here, we see that the data have been divided nearly evenly among four geographic regions

Explore relationships among features – the correlation matrix

- Before fitting a regression model to data, it can be useful to determine how the independent variables are related to the dependent variable and each other.
- A correlation matrix provides a quick overview of these relationships

Exploring relationships among features – the correlation matrix

- There are various rules of thumb used to interpret correlation strength.
- One method assigns a status of “weak” to values between 0.1 and 0.3, “moderate” to the range of 0.3 to 0.5, and “strong” to values above 0.5 (these also apply to similar ranges of negative correlations).
- However, these thresholds may be too lax for some purposes. Often, the correlation must be interpreted in context.
- For data involving human beings, a correlation of 0.5 may be considered extremely high, while for data generated by mechanical processes, a correlation of 0.5 may be weak.

Correlation matrix

To create a correlation matrix for the four numeric variables in the insurance data frame, use the **cor()** command:

```
> cor(insurance[c("age", "bmi", "children", "charges")])
               age       bmi   children    charges
age      1.0000000 0.1092719 0.04246900 0.29900819
bmi      0.1092719 1.0000000 0.01275890 0.19834097
children 0.0424690 0.0127589 1.00000000 0.06799823
charges  0.2990082 0.1983410 0.06799823 1.00000000
```

None of the correlations in the matrix are considered strong, but there are some notable associations.

- age and charges
- bmi and charge
- children and charges.
- age and bmi

Visualize relationships among features – the scatterplot matrix

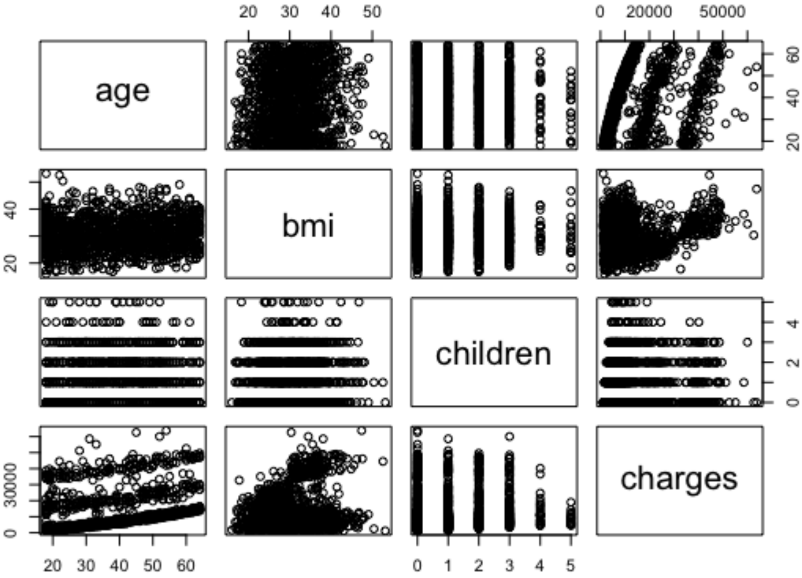
- An alternative is to create a scatterplot matrix.
- A scatterplot matrix simply a collection of scatterplots arranged in a grid.
- It is used to detect patterns among three or more variables

Visualizing relationships among features – the scatterplot matrix

pairs() function is provided in a default R installation and provides basic functionality for producing scatterplot matrices.

```
pairs(insurance[c("age", "bmi", "children", "charges")])
```

The scatterplot matrix



The scatterplot matrix

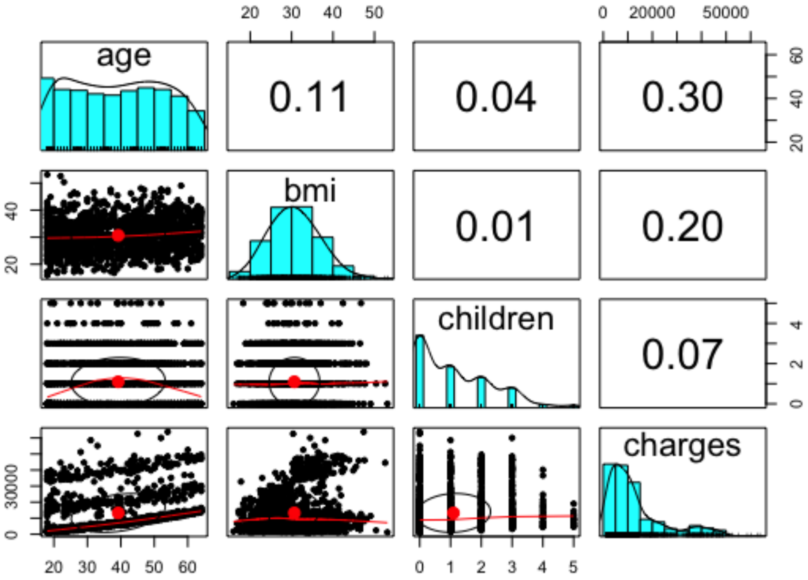
- Although some look like random clouds of points, a few seem to display some trends.
- The relationship between **age** and **charges** displays several relatively straight lines,
- **bmi** and **charges** has two distinct groups of points.

The scatterplot matrix

An enhanced scatterplot matrix can be created with the **pairs.panels()** function in the psych package.

```
> install.packages("psych")

#load it to your work session
> library (psych)

#create a scatter plot
>pairs.panels(insurance[c("age", "bmi", "children", "charges")])
```

The scatterplot matrix



The scatterplot matrix

- Above the diagonal, the scatterplots have been replaced with a correlation matrix.
- On the diagonal, a histogram depicting the distribution of values for each feature is shown.
- Below the diagonal now are presented with additional visual information.

The scatterplot matrix

- The correlation between the two variables is indicated by the shape of the ellipse; the more it is stretched, the stronger the correlation.
- An almost perfectly round oval, as with bmi and children, indicates a very weak correlation (in this case 0.01).
- The dot at the center of the ellipse indicates the point of the mean value for the x axis variable and y axis variable.

Create a model

```
ins_model <- lm(charges ~ age + children + bmi + gender + smoker
    + region, data = insurance)
```

View the model

```
 > ins_model

Call:
lm(formula = charges ~ age + children + bmi + gender + smoker +
    region, data = insurance)

Coefficients:
    (Intercept)                age           children                bmi
       -11938.5              256.9              475.5              339.2
     gendermale          smokeryes    regionnorthwest    regionsoutheast
         -131.3            23848.5             -353.0            -1035.0
regionsouthwest
         -960.1
```

Interpretation of the model

The estimated coefficients

- For instance, for each year that age increases, we would expect $256.90 higher medical expenses on average, assuming everything else is equal.
- Similarly, each additional child results in an average of $475.50 in additional medical expenses each year
- Each unit of BMI increase is associated with an increase of $339.20 in yearly medical costs.

```
> ins_model$coefficients
    (Intercept)         age    children         bmi     gendermale    smokeryes
    -11938.5386 256.8564    475.5005   339.1935     -131.3144   23848.5345
regionnorthwest regionsoutheast regionsouthwest
      -352.9639      -1035.0220       -960.0510
```

Factor type

- Here we only specified six features in our model formula
- However, there are eight coefficients reported in addition to the intercept.
- The **lm()** function automatically applied a technique known as dummy coding to each of the **factor type** variables we included in the model.

Factor type: dummy coding

- Dummy coding allows a nominal feature to be treated as numeric by creating a binary variable for each category of the feature, which is set to 1 if the observation falls into that category or 0 otherwise.
- For example, the **gender** variable has two categories, male and female. This will be split into two binary values,
  - gendermale
  - genderfemale
- For four-category feature **region**, it is split into four variables
  - regionnorthwest
  - regionsoutheast
  - regionsouthwest
  - regionnortheast

Dummy coding: the reference category

- When adding a dummy-coded variable to a regression model, one category is always left out to serve as the reference category.
- The estimates are then interpreted relative to the reference.
- In our model, R automatically held out the **genderfemale**, **smokerno**, and **regionnortheast** variables, making female non-smokers in the northeast region the **reference group**.

Dummy coding: the reference category

```
> round(ins_model$coefficients, 2)
    (Intercept)          age     children          bmi   gendermale    smokeryes
      -11938.54       256.86       475.50       339.19      -131.31     23848.53
regionnorthwest regionsoutheast regionsouthwest
        -352.96        -1035.02         -960.05
```

- Males have \$131.30 less medical costs each year relative to females
- Smokers cost an average of \$23,848.50 more than non-smokers.
- The coefficient for each of the other three regions in the model is negative, which implies that the northeast region tends to have the highest average medical expenses.

Evaluating model performance

```
> summary (ins_model)
Call:
lm(formula = charges ~ age + children + bmi + gender + smoker +
    region, data = insurance)
Residuals:
     Min       1Q   Median       3Q      Max
-11304.9  -2848.1   -982.1   1393.9  29992.8
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
age                256.9       11.9  21.587  < 2e-16 ***
children           475.5      137.8   3.451 0.000577 ***
bmi                339.2       28.6  11.860  < 2e-16 ***
gendermale        -131.3      332.9  -0.394 0.693348
smokeryes        23848.5      413.1  57.723  < 2e-16 ***
regionnorthwest   -353.0      476.3  -0.741 0.458769
regionsoutheast  -1035.0      478.7  -2.162 0.030782 *
regionsouthwest   -960.0      477.9  -2.009 0.044765 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 6062 on 1329 degrees of freedom
Multiple R-squared:  0.7509,Adjusted R-squared:  0.7494
F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

Evaluating model performance: Residuals

The Residuals section provides summary statistics for the errors in our predictions

- Residual is equal to the true value minus the predicted value
- The maximum error of 29992.8 suggests that the model under-predicted expenses by nearly $30,000 for at least one observation.
- On the other hand, 50% of errors fall within the 1Q and 3Q values (the first and third quartile), so the majority of predictions were between $2,850 over the true value and $1,400 under the true value.

```
> summary(ins_model$residuals)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-11304.9  -2848.1   -982.1      0.0   1393.9  29992.8
```

Evaluating model performance

- Since the **R-squared** value is 0.7494, we know that nearly 75 percent of the variation in the dependent variable is explained by our model.
- As models with more features always explain more variation, the **Adjusted R-squared** value corrects R-squared by penalizing models with a large number of independent variables.
  - It is useful for comparing the performance of models with different numbers of explanatory variables.
- Overall p value on the basis of F−statistic, often p value less than 0.05 indicate that overall model is significant

Improve model

- The effect of age on medical expenditures may not be constant throughout all age values; the treatment may become disproportionately expensive for the oldest populations.
- To account for a non-linear relationship, we can add a higher order term to the regression model, treating the model as a polynomial. In effect, we will be modeling a relationship like this:

$$y = \alpha + \beta_1 x + \beta_2 x^2 \tag{6}$$

- To add the non-linear age to the model, we simply need to create a new variable:

```
> insurance$age2 <- insurance$age^2
```

Improve model: converting a numeric variable to a binary indicator

- Suppose we have a hunch that the effect of a feature is not cumulative, but rather it has an effect only once a specific threshold has been reached.
- For instance, BMI may have zero impact on medical expenditures for individuals in the normal weight range, but it may be strongly related to higher costs for the obese (that is, BMI of 30 or above).

```
#For BMI greater than or equal to 30, we will return 1,
    otherwise 0:
insurance$bmi30 <- ifelse(insurance$bmi >= 30, 1, 0)
```

Improve model: adding interaction effects

- So far, we have only considered each feature's individual contribution to the outcome.
- What if certain features have a combined impact on the dependent variable?
- For instance, smoking and obesity may have harmful effects separately, but it is reasonable to assume that their combined effect may be worse than the sum of each one alone
- The * operator is shorthand that instructs R to model

```
charges ~ bmi30*smoker

# the * operator is shorthand that instructs R to model
charges ~ bmi30 + smokeryes + bmi30:smokeryes
```

The : (colon) operator in the expanded form indicates that **bmi30:smokeryes** is the interaction between the two variables.

New model:

```
> ins_model2 <- lm(charges ~ age + age2 + children + bmi + gender
    + bmi30*smoker + region, data = insurance)
```

Model performance

```
> summary (ins_model2)
Call:
lm(formula = charges ~ age + age2 + children + bmi + gender +
    bmi30 * smoker + region, data = insurance)
Residuals:
     Min       1Q   Median       3Q      Max
-17296.4  -1656.0  -1263.3   -722.1  24160.2
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       134.2509  1362.7511   0.099 0.921539
age               -32.6851    59.8242  -0.546 0.584915
age2                3.7316     0.7463   5.000 6.50e-07 ***
children          678.5612   105.8831   6.409 2.04e-10 ***
bmi               120.0196    34.2660   3.503 0.000476 ***
gendermale       -496.8245   244.3659  -2.033 0.042240 *
bmi30           -1000.1403   422.8402  -2.365 0.018159 *
smokeryes       13404.6866   439.9491  30.469  < 2e-16 ***
regionnorthwest  -279.2038   349.2746  -0.799 0.424212
regionsoutheast  -828.5467   351.6352  -2.356 0.018604 *
regionsouthwest -1222.6437   350.5285  -3.488 0.000503 ***
bmi30:smokeryes 19810.7533   604.6567  32.764  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 4445 on 1326 degrees of freedom
Multiple R-squared:  0.8664,Adjusted R-squared:  0.8653
F-statistic: 781.7 on 11 and 1326 DF,  p-value: < 2.2e-16
```

Performance comparison

```
#Save summary of the regression model to a object
> perf_ins_model <- summary(ins_model)
> perf_ins_model2 <- summary(ins_model2)

#Obtain the attribute list of the object perf_ins_model
> attributes(perf_ins_model)
$names
 [1] "call"  "terms"  "residuals"  "coefficients"  "aliased"
 [6] "sigma"  "df"  "r.squared"  "adj.r.squared"  "fstatistic"
[11] "cov.unscaled"

$class
[1] "summary.lm"

#Extract R square and adjust R square values of the two models and save it to a data frame
> model_comp <- data.frame(r.square=perf_ins_model$r.squared, adj.r.squared=perf_ins_
    model$adj.r.squared)
> model_comp <- rbind(model_comp, c(r.square=perf_ins_model2$r.squared, adj.r.squared=
    perf_ins_model2$adj.r.squared ))
> rownames(model_comp) = c("model1", "model2")
```

Performance comparison

```
#If the local library does not have DT package, install it
> if(!require(DT)) install.packages("DT")
> library(DT)

> datatable(round(model_comp, 3))
```

Show 10 entries Search:

	r.square	adj.r.squared
model1	0.751	0.749
model2	0.866	0.865

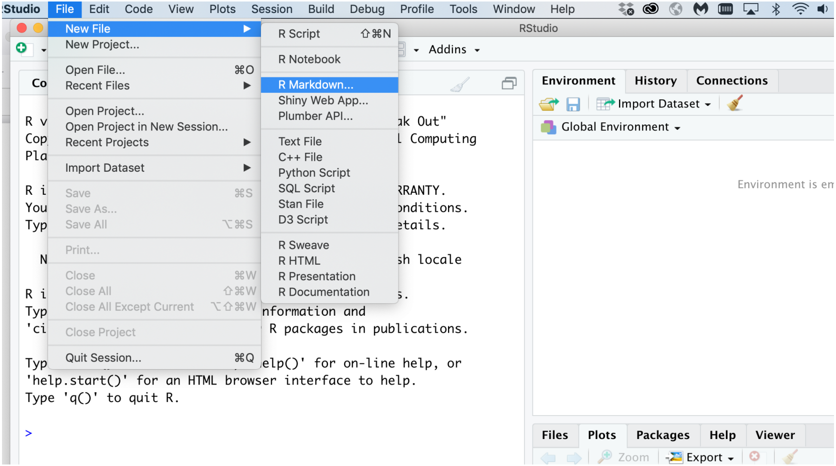
Showing 1 to 2 of 2 entries Previous 1 Next

R Makrdown

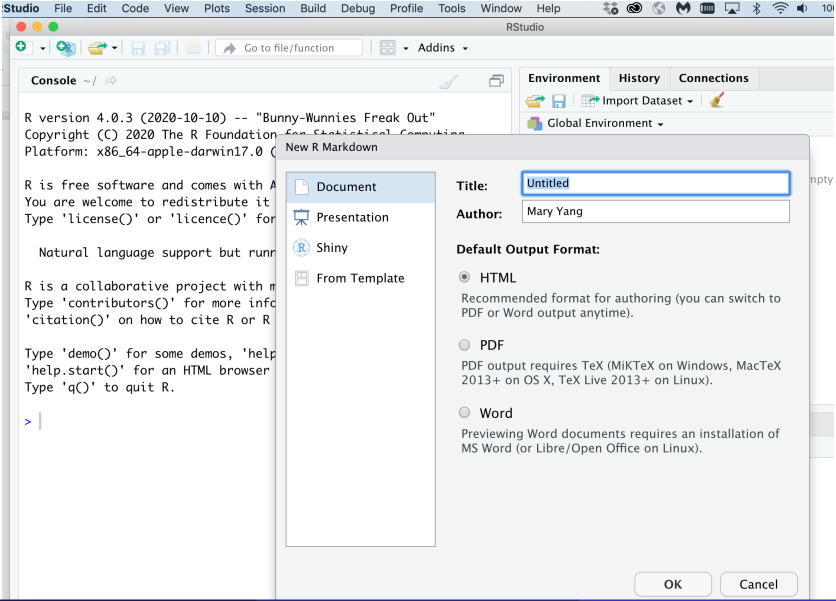
- R Markdown provides an authoring framework for data science. You can use a single R Markdown file to both
  - save and execute code
  - generate high quality reports that can be shared with an audience
- R Markdown documents are fully reproducible.
- R Markdown supports dozens of static and dynamic output formats including HTML, PDF, MS Word, Beamer, Tufte-style handouts, books, dashboards, shiny

```
> install.packages("rmarkdown")

>libary(rmarkdown) # load the package into your workspace
```

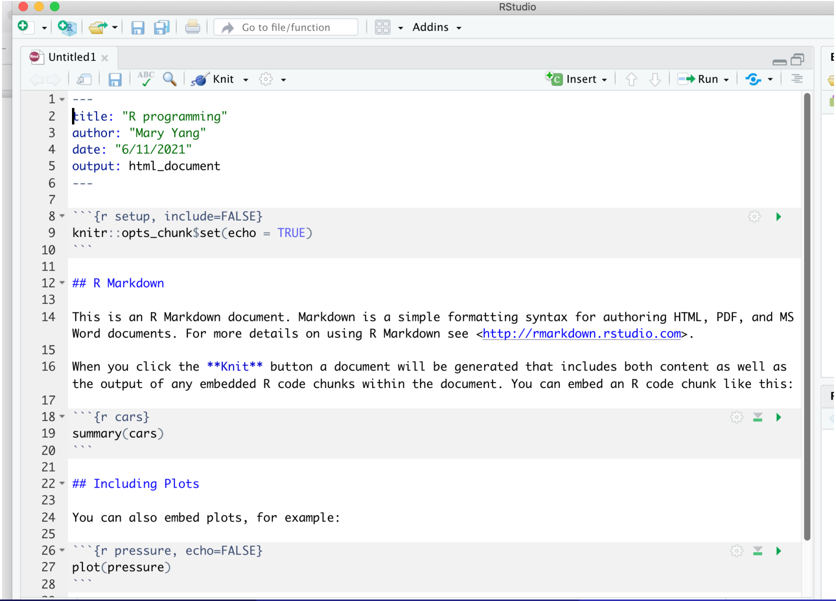
R Makrdown


RStudio
File
Edit
Code
View
Plots
Session
Build
Debug
Profile
Tools
Window
Help
New File
New Project...
Open File... ⌘O
Recent Files
Open Project...
Open Project in New Session...
Recent Projects
Import Dataset
Save ⌘S
Save As...
Save All ⌥⌘S
Print...
Close ⌘W
Close All ⇧⌘W
Close All Except Current ⌥⇧⌘W
Close Project
Quit Session... ⌘Q
R Script ⇧⌘N
R Notebook
R Markdown...
Shiny Web App...
Plumber API...
Text File
C++ File
Python Script
SQL Script
Stan File
D3 Script
R Sweave
R HTML
R Presentation
R Documentation
Addins
Environment
History
Connections
Import Dataset
Global Environment
Environment is em
Files
Plots
Packages
Help
Viewer
Zoom
Export
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

R Makrdown

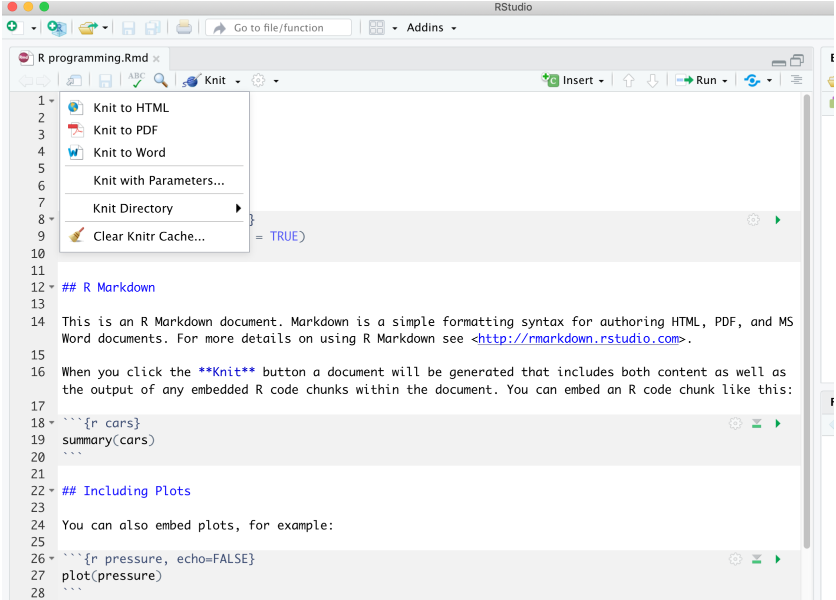

RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
RStudio
Go to file/function
Addins
Console ~/
Environment History Connections
Import Dataset
Global Environment
R version 4.0.3 (2020-10-10) -- "Bunny-Wunnies Freak Out"
Copyright (C) 2020 The R Foundation
Platform: x86_64-apple-darwin17.0
R is free software and comes with
You are welcome to redistribute it
Type 'license()' or 'licence()' for
Natural language support but runn
R is a collaborative project with
Type 'contributors()' for more info
'citation()' on how to cite R or R
Type 'demo()' for some demos, 'help
'help.start()' for an HTML browser
Type 'q()' to quit R.
New R Markdown
Document
Presentation
Shiny
From Template
Title: Untitled
Author: Mary Yang
Default Output Format:
HTML
Recommended format for authoring (you can switch to PDF or Word output anytime).
PDF
PDF output requires TeX (MiKTeX on Windows, MacTeX 2013+ on OS X, TeX Live 2013+ on Linux).
Word
Previewing Word documents requires an installation of MS Word (or Libre/Open Office on Linux).
OK
Cancel

R Makrdown

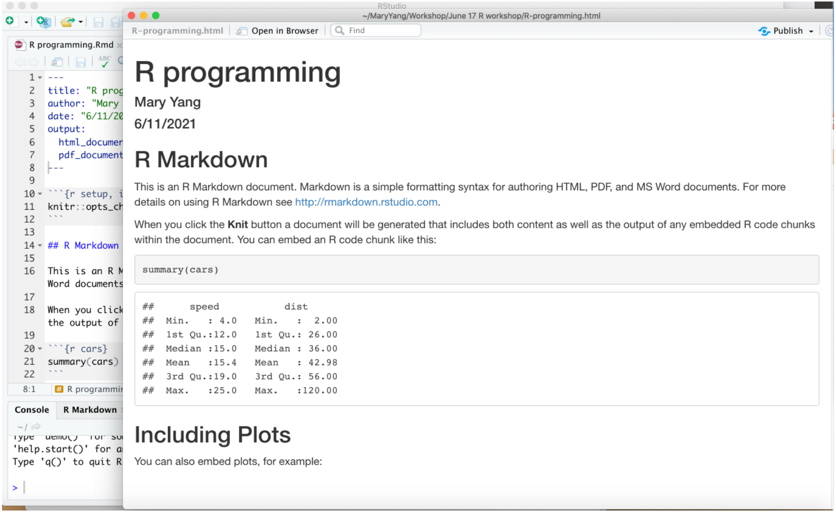

RStudio
Go to file/function
Addins
Untitled1
Knit
Insert
Run
---
title: "R programming"
author: "Mary Yang"
date: "6/11/2021"
output: html_document

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```
R Markdown
This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:
```{r cars}
summary(cars)
```
Including Plots
You can also embed plots, for example:
```{r pressure, echo=FALSE}
plot(pressure)
```

R Makrdown


RStudio
Go to file/function
Addins
R programming.Rmd
Knit
Insert
Run
Knit to HTML
Knit to PDF
Knit to Word
Knit with Parameters...
Knit Directory
Clear Knitr Cache...
}
= TRUE)
R Markdown
This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:
```{r cars}
summary(cars)
```
Including Plots
You can also embed plots, for example:
```{r pressure, echo=FALSE}
plot(pressure)
```

R Makrdown


RStudio
~/MaryYang/Workshop/June 17 R workshop/R-programming.html
R-programming.html
Open in Browser
Find
Publish
R programming.Rmd
---
title: "R prog
author: "Mary
date: "6/11/20
output:
html_documen
pdf_document

```{r setup, i
knitr::opts_ch
```
R Markdown
This is an R M
Word documents
When you click
the output of
```{r cars}
summary(cars)
```
8:1
R programmi
Console
R Markdown
'help.start()' for a
Type 'q()' to quit R
R programming
Mary Yang
6/11/2021
R Markdown
This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.
When you click the Knit button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:
summary(cars)
speed   dist
Min.   : 4.0   Min.   : 2.00
1st Qu.:12.0   1st Qu.: 26.00
Median :15.0   Median : 36.00
Mean   :15.4   Mean   : 42.98
3rd Qu.:19.0   3rd Qu.: 56.00
Max.   :25.0   Max.   :120.00
Including Plots
You can also embed plots, for example:

header

Syntax	Becomes
`# Header 1`	Header 1
`## Header 2`	Header 2
`### Header 3`	Header 3
`#### Header 4`	Header 4
`##### Header 5`	Header 5
`###### Header 6`	Header 6

Code chunk

Syntax	Becomes
Make a code chunk with three back ticks followed by an r in braces. End the chunk with three back ticks: ` ```{r} ` `paste("Hello", "World!")` ` ``` `	Make a code chunk with three back ticks followed by an r in braces. End the chunk with three back ticks: `paste("Hello", "World!")` `## [1] "Hello World!"`

Code chunk

```
Add chunk options within braces. For example,
`echo=FALSE` will prevent source code from being
displayed:

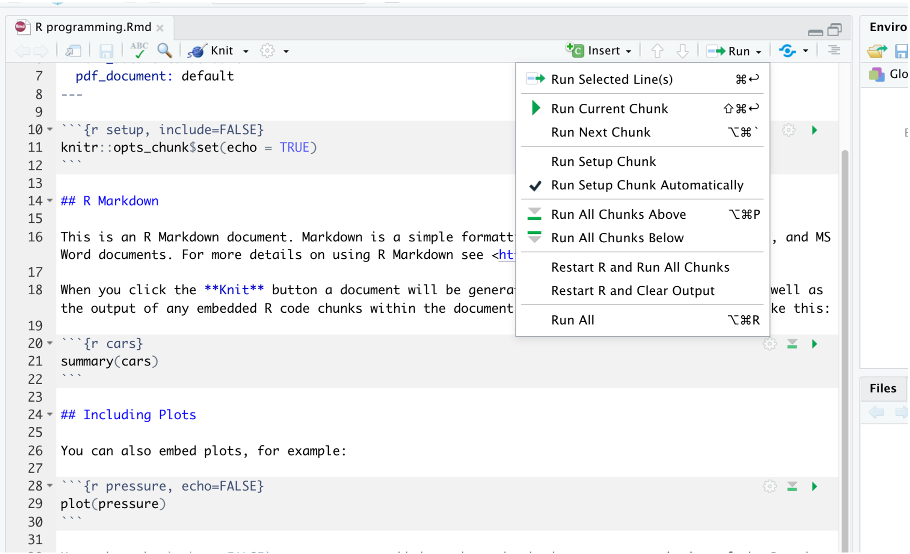
```{r eval=TRUE, echo=FALSE}
paste("Hello", "World!")
```
```

Add chunk options within braces. For example, `echo=FALSE` will prevent source code from being displayed:

```
## [1] "Hello World!"
```

See R Markdown Reference Guide for a complete list of knitr chunk options.

Execute code



Reference

- R Graphics Cookbook by *Winston Wang*
- Using R for Data Analysis and Graphics by *J H Maindonald*
- The Art of R Programming by *NormanMatolff*
- Machine Learning with R by *BrettLantz*

Exercises

Exercise 1:

- Bulid a linear regression model tp predict the sale using facebook advertisement budget
- Plot a linear regression line to fit the sale data
- Is this a good model?

Exerise 2:

- Draw a boxplot of medical charges in different regions.