In the Name of God

Sharif University of Technology
Department of Computer Engineering

Engineering Probabilistics

Winter 1401

EDA (Exploratory Data Analysis)

Ali Mehrabani, Sina Imani

Amirreza Azari - 99101087

"You see, but you do not observe. The distinction is clear." - Sherlock Holmes

Sir Arthur Conan Doyle 1859-1930
from *The Scandal of Bohemia*

Introduction

Imagine that you have gone to the doctor's because you feel ill. The doctor is equipped with lots of theoretical knowledge about different kinds of illness. She/He also has access to a wide variety of powerful medical tools invented to help ill people, like chemical solutions and tablets. But before using these tools, it is necessary to perform a **diagnosis** step first.

Similar statements are true when talking about statistical inference. Over the years, many powerful probabilistic techniques have been developed to help gain useful information from data. For example:

Hypothesis testing provides a systematic process in which one can check if there is a 'meaningful' distinction between two groups of observations, each having a certain property. You can, for example, test if male employees are paid higher amounts on average than female employees in a country, using a relatively small sample of male and female employees.

Linear Regression methods are designed to find linear relations between *variables*, which are properties of collected datapoints. Using these methods, you can build a *model* that takes the collected data and estimates the desired property of a new data point. Estimating the price of a car given its technical features could be one usual application of such methods.

The list continues even more. But the important point to keep in mind here is that none of these *automated* techniques can give you a deep and complete understanding of the data on their own. In fact, some prior insight into data is *required* to be able to use such systematic procedures appropriately. As the doctor needs diagnosis to select the appropriate medical care - say, a drug - and select the appropriate way of using it - say, its dose - we need to examine our data before using *automated* techniques.

This is when 'Exploratory Data Analysis' comes in. Loosely speaking, EDA is the process of 'diagnosing' the dataset. You first try to gain some insight into data before further analyses. This can be useful in many ways.

- Unexpected discoveries about data, which can lead to further studies
- Suggesting appropriate hypotheses to test
- Support the selection of appropriate statistical tools and techniques
- Being more familiar with data and so getting able to analyze it both intuitively and systematically

So, How to Perform EDA?

As mentioned before, EDA is the process of 'diagnosing' and getting insight into data. So it is more than anything, a state of mind and does not have any strict 'set of rules'. Although there are guidelines about which EDA techniques are useful in what circumstances, there is an important degree of looseness and art to EDA. Competence and confidence come with practice, experience, and close observation of others. Also, EDA need not be restricted to techniques you have seen before; **sometimes you need to invent a new way of looking at your data**.

Having this said, there are some 'conventions' in EDA:

- Take visualization seriously! People are not very good at looking at a column of numbers or a whole spreadsheet and then determining important characteristics of the data. They find looking at numbers to be tedious, boring, and/or overwhelming. Exploratory data analysis techniques have been devised as an aid in this situation.
- Be eagle-eyed! Pay attention to what you are seeing. Does the data meet your expectations? Are there any anomalies? What are the most common values of variables, and why? During EDA you should be careful and ask intelligent questions about data. You:
  1. Generate questions about your data.
  2. Search for answers by visualizing, transforming, and modeling your data.
  3. Use what you learn to refine your questions and/or generate new questions
- Tip: EDA usually starts with considering the distribution of variables, and correlations between them. Paying attention to the form of distributions is important.

Remember, no technology can replace that precise, rational thinking of humans that reveals the hidden aspects of mysterious problems! To be a data scientist is similar to being a detective. They both pay attention to the smallest details, put the clues in order, and try to see the underlying reasons for each property.

An Example

Findings from EDA are orthogonal to the primary analysis task. To illustrate, consider an example from Cook et al. where the analysis task is to find the variables which best predict the tip that a dining party will give to the waiter. Here is the original source of this example. (https://en.wikipedia.org/wiki/Exploratory_data_analysis#cite_note-12) The variables available in the data collected for this task are: the tip amount, total bill, payer gender, smoking/non-smoking section, time of day, day of the week, and size of the party. The primary analysis task is approached by fitting a regression model where the tip rate is the response variable. The fitted model is

(tip rate) = 0.18 - 0.01 × (party size) which says that as the size of the dining party increases by one person (leading to a higher bill), the tip rate will decrease by 1%, on average.

However, exploring the data reveals other interesting features not described by this model.

Histogram of tip amounts where the bins cover $1 increments. The distribution of values is skewed right and uni-modal (= having one mode), as is common in distributions of small, non-negative quantities.

Histogram of tip amounts where the bins cover $0.10 increments. An interesting phenomenon is visible: peaks occur at the whole-dollar and half-dollar amounts, which is caused by customers picking round numbers as tips. This behavior is common to other types of purchases too, like gasoline.

Scatterplot of tips vs. bill. Points below the line correspond to tips that are lower than expected (for that bill amount), and points above the line are higher than expected. We might expect to see a tight, positive linear association, but instead, see a variation that increases with tip amount. In particular, there are more points far away from the line in the lower right than in the upper left, indicating that more customers are very cheap than very generous.

Scatterplot of tips vs. bill separated by payer gender and smoking section status. Smoking parties have a lot more variability in the tips that they give. Males tend to pay the (few) higher bills, and the female non-smokers tend to be very consistent tippers (with three conspicuous exceptions shown in the sample).

What is learned from the plots is different from what is illustrated by the regression model, even though the experiment was not designed to investigate any of these other trends. The patterns found by exploring the data suggest hypotheses about tipping that may not have been anticipated in advance, and which could lead to interesting follow-up experiments where the hypotheses are formally stated and tested by collecting new data.

Let's Start!

Now that you have been introduced to the concept of EDA, it's time for trying it in practice. In this assignment, two questions have been designed. In the first one, we want to practice the skills needed for a good EDA. For doing so, we complete an EDA process step by step. This is like performing pre-defined tasks instead of doing free research - which is less appealing but is more beneficial at the start. We use the 'penguins.csv' dataset in this question, which is a very appropriate one for practicing EDA. This dataset contains information about more than 300 penguins that live in the Palmer Archipelago, located northeast of Antarctica. This information is real and potentially contains lots of knowledge about these penguins.

In the second question, you pick a favorite dataset and do EDA on it, without limitations!

When doing this exercise, you may find some questions a bit ambiguous or vague. You may not also be confident about the correctness of your answers. Don't panic! There is no 'right answer' in this exercise. The important thing to us is your analyzing abilities and your advancements in *diagnosing* the data. For example, when we require you to predict the correlation between two random variables, we are **, not** interested to check if your prediction was correct or not. Rather, we are really interested to check if you've performed meaningful reasoning and analysis after seeing that your prediction was correct - or incorrect.

It is highly recommended to use some appropriate environment - like RStudio - to write your code and consider your plots, rather than the browser you're reading this text on.

Copyright Notice

Palmer Archipelago (Antarctica) penguin data Citation:

Gorman KB, Williams TD, Fraser WR (2014) Ecological Sexual Dimorphism and Environmental Variability within a Community of Antarctic Penguins (Genus Pygoscelis). PLoS ONE 9(3): e90081. doi:10.1371/journal.pone.0090081

References and Further Reading

These contents have been used in order to write this article:

- https://en.wikipedia.org/wiki/Exploratory_data_analysis (https://en.wikipedia.org/wiki/Exploratory_data_analysis) (mainly for the example EDA). The example is originally introduced in Cook, D. and Swayne, D.F. (with A. Buja, D. Temple Lang, H. Hofmann, H. Wickham, M. Lawrence) (2007) "Interactive and Dynamic Graphics for Data Analysis: With R and GGobi" Springer, 978-0387717616
- Experimental Design and Analysis - By Prof. Howard J. Seltman, Carnegie Mellon University (mainly for describing EDA). The whole book is available at https://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf (https://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf). Chapter 4 is related to EDA: https://www.stat.cmu.edu/~hseltman/309/Book/chapter4.pdf (https://www.stat.cmu.edu/~hseltman/309/Book/chapter4.pdf)
- https://www.bbc.com/future/article/20160107-what-sherlock-holmes-tells-us-about-the-mind (https://www.bbc.com/future/article/20160107-what-sherlock-holmes-tells-us-about-the-mind) (mainly for emphasizing on the importance of paying attention to clues)
- https://r4ds.had.co.nz/exploratory-data-analysis.html#exploratory-data-analysis (https://r4ds.had.co.nz/exploratory-data-analysis.html#exploratory-data-analysis) (mainly for describing EDA). It is an excellent source for further reading.

Question 1 - Palmer Penguins

In this question we have access to the following information of more than 300 penguins: specie, island, bill length, bill depth, flipper length, body mass and sex. Bill length and depth are measured as below.

Getting Familiar with Data / Non-graphical EDA

For start, we load the attached .csv file as a dataframe:

In [5]:
```
df <- read.csv("penguins.csv")
head(df)
```

A data.frame: 6 × 7

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
	<chr>	<chr>	<dbl>	<dbl>	<int>	<int>	<chr>
1	Adelie	Torgersen	39.1	18.7	181	3750	MALE
2	Adelie	Torgersen	39.5	17.4	186	3800	FEMALE
3	Adelie	Torgersen	40.3	18.0	195	3250	FEMALE
4	Adelie	Torgersen	NA	NA	NA	NA	NA
5	Adelie	Torgersen	36.7	19.3	193	3450	FEMALE
6	Adelie	Torgersen	39.3	20.6	190	3650	MALE

Call the 'summary' function on df and pay attention to its result. As you see, some columns contain numerical data and some contain characters. Try to gain some sense of the distribution of numerical columns, given their min, max, and quartiles.

In [2]:
```
summary(df)
```

```
   species             island          bill_length_mm  bill_depth_mm  
 Length:344         Length:344         Min.   :32.10   Min.   :13.10  
 Class :character   Class :character   1st Qu.:39.23   1st Qu.:15.60  
 Mode  :character   Mode  :character   Median :44.45   Median :17.30  
                                       Mean   :43.92   Mean   :17.15  
                                       3rd Qu.:48.50   3rd Qu.:18.70  
                                       Max.   :59.60   Max.   :21.50  
                                       NA's   :2       NA's   :2      
 flipper_length_mm  body_mass_g       sex           
 Min.   :172.0     Min.   :2700   Length:344        
 1st Qu.:190.0     1st Qu.:3550   Class :character  
 Median :197.0     Median :4050   Mode  :character  
 Mean   :200.9     Mean   :4202                     
 3rd Qu.:213.0     3rd Qu.:4750                     
 Max.   :231.0     Max.   :6300                     
 NA's   :2         NA's   :2                        
```

Suppose that in some distribution, the mean is greater than the median. What does this say about the distribution? What if the mean is less than, or equal to the median?
The mean, mode and median can be used to figure out if you have a positively or negatively skewed distribution. If the mean is greater than the mode, the distribution is positively skewed. If the mean is less than the mode, the distribution is negatively skewed. If the mean is greater than the median, the distribution is positively skewed. If the mean is less than the median, the distribution is negatively skewed. When the mean is greater than the median, the shape of the distribution is skewed to the right. This means that the bulk of the data are concentrated on the left and there is a long tail stretching to the right.

The 'table' function takes an array as input and outputs its frequency table. Which columns of df are better to call this function on? Do the call for them.

In [3]:
```
table(df$species)
table(df$island)
table(df$sex)
```

```
   Adelie Chinstrap    Gentoo 
      152        68       124 


   Biscoe     Dream Torgersen 
      168       124        52 


     . FEMALE   MALE 
     1    165    168 
```

For initial exposure to data, one can use more advanced functions rather than built-in R functions. The function 'skim' in the library 'skimr' is one such powerful function. Call this function on df and consider the result carefully. If you have not installed 'skimr', first do so by issuing the command below:

install.packages("skimr").

In [5]:
```
#install.packages("skimr")
library(skimr)
skim(df)
```

```
── Data Summary ────────────────────────
                           Values
Name                       df
Number of rows             344
Number of columns          7
_________________________
Column type frequency:
  character                3
  numeric                  4
_________________________
Group variables            None

── Variable type: character ────────────────────────────────────────────────────
  skim_variable n_missing complete_rate min max empty n_unique whitespace
1 species               0         1       6   9     0        3          0
2 island                0         1       5   9     0        3          0
3 sex                  10         0.971   1   6     0        3          0

── Variable type: numeric ──────────────────────────────────────────────────────
  skim_variable     n_missing complete_rate   mean     sd     p0    p25    p50
1 bill_length_mm            2         0.994   43.9   5.46   32.1   39.2   44.4
2 bill_depth_mm             2         0.994   17.2   1.97   13.1   15.6   17.3
3 flipper_length_mm         2         0.994  201.   14.1   172    190    197
4 body_mass_g               2         0.994 4202.  802.   2700   3550   4050
     p75   p100 hist
1   48.5   59.6 ▃▇▇▆▁
2   18.7   21.5 ▃▇▇▇▃
3  213    231   ▃▇▁▆▂
4 4750   6300   ▃▇▆▃▂
```

```
Warning message in is.null(text_repr) || nchar(text_repr) == 0L:
"'length(x) = 17 > 1' in coercion to 'logical(1)'"
```

A skim_df: 7 × 17

	skim_type	skim_variable	n_missing	complete_rate	character.min	character.max	character.empty	character.n_unique	character.whitespace	numeric.mean	numeric.sd	numeric.p0	num
	<chr>	<chr>	<int>	<dbl>	<int>	<int>	<int>	<int>	<int>	<dbl>	<dbl>	<dbl>	
1	character	species	0	1.0000000	6	9	0	3	0	NA	NA	NA	
2	character	island	0	1.0000000	5	9	0	3	0	NA	NA	NA	
3	character	sex	10	0.9709302	1	6	0	3	0	NA	NA	NA	
4	numeric	bill_length_mm	2	0.9941860	NA	NA	NA	NA	NA	43.92193	5.459584	32.1	
5	numeric	bill_depth_mm	2	0.9941860	NA	NA	NA	NA	NA	17.15117	1.974793	13.1	
6	numeric	flipper_length_mm	2	0.9941860	NA	NA	NA	NA	NA	200.91520	14.061714	172.0	
7	numeric	body_mass_g	2	0.9941860	NA	NA	NA	NA	NA	4201.75439	801.954536	2700.0	

As you can see, some rows of data contain values that are NA. NA stands for Not Available. When encountering such rows, we can follow different strategies. One is to replace them with specific values, like mean or median. In this case, due to the relatively small number of NAs, we can simply remove them from our dataframe. So write a code to remove all rows in df that contain NA values.

In [8]:
```
new_df <- na.omit(df)
head(new_df, 20)
```

A data.frame: 20 × 7

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
	<chr>	<chr>	<dbl>	<dbl>	<int>	<int>	<chr>
1	Adelie	Torgersen	39.1	18.7	181	3750	MALE
2	Adelie	Torgersen	39.5	17.4	186	3800	FEMALE
3	Adelie	Torgersen	40.3	18.0	195	3250	FEMALE
5	Adelie	Torgersen	36.7	19.3	193	3450	FEMALE
6	Adelie	Torgersen	39.3	20.6	190	3650	MALE
7	Adelie	Torgersen	38.9	17.8	181	3625	FEMALE
8	Adelie	Torgersen	39.2	19.6	195	4675	MALE
13	Adelie	Torgersen	41.1	17.6	182	3200	FEMALE
14	Adelie	Torgersen	38.6	21.2	191	3800	MALE
15	Adelie	Torgersen	34.6	21.1	198	4400	MALE
16	Adelie	Torgersen	36.6	17.8	185	3700	FEMALE
17	Adelie	Torgersen	38.7	19.0	195	3450	FEMALE
18	Adelie	Torgersen	42.5	20.7	197	4500	MALE
19	Adelie	Torgersen	34.4	18.4	184	3325	FEMALE
20	Adelie	Torgersen	46.0	21.5	194	4200	MALE
21	Adelie	Biscoe	37.8	18.3	174	3400	FEMALE
22	Adelie	Biscoe	37.7	18.7	180	3600	MALE
23	Adelie	Biscoe	35.9	19.2	189	3800	FEMALE
24	Adelie	Biscoe	38.2	18.1	185	3950	MALE
25	Adelie	Biscoe	38.8	17.2	180	3800	MALE

Graphical EDA

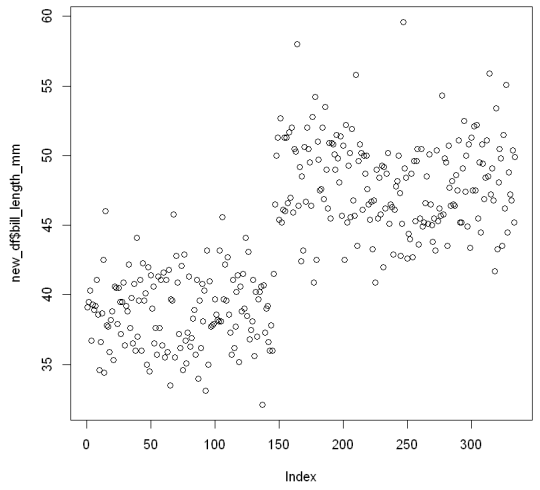
As mentioned before, performing EDA is mainly done by different means of visualization. In this section, we are going to practice some of these techniques. We also practice the 'cyclic' process of EDA: Examining data, Generating questions, Examining the data again to find answers, Refining the questions or generating new questions, and so on!

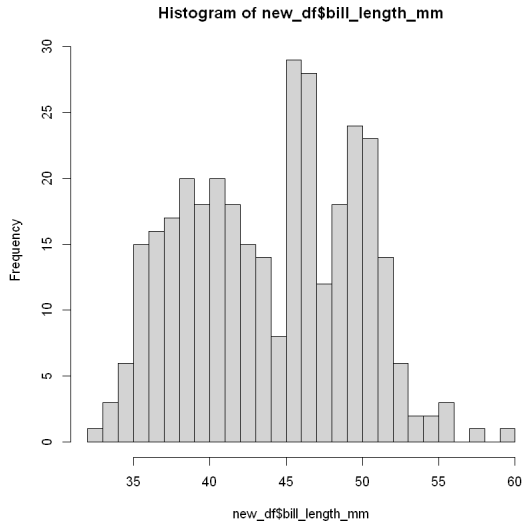
We start by considering the distribution of single variables. Write a code to draw a plot, showing the distribution of the 'bill_length' variable.

What distribution do you think the 'bill_length' variable has? Write your opinion here, before running the code.

Your Prediction of Distribution: Maybe normal

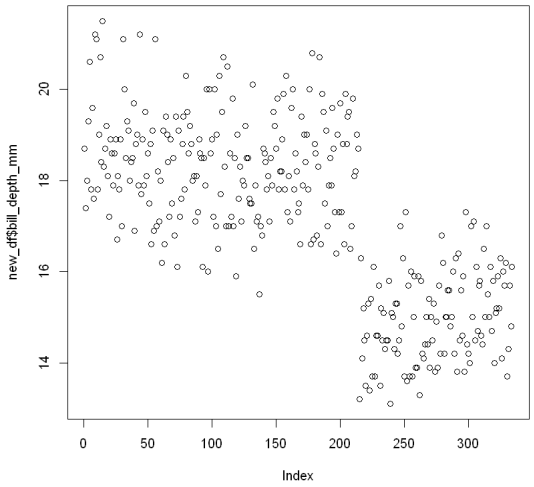
In [7]:
```
plot(new_df$bill_length_mm)
hist(new_df$bill_length_mm, breaks = 25)
```

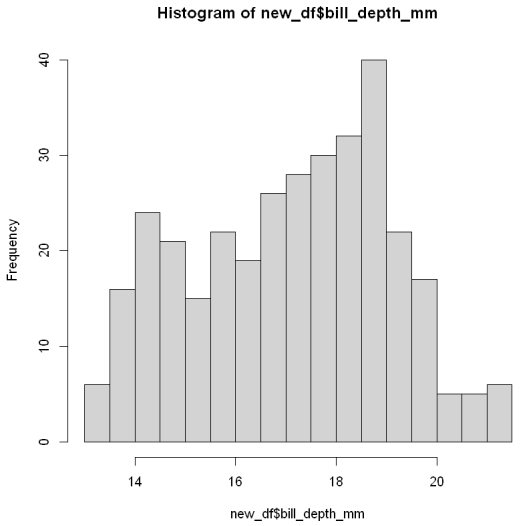




Let's repeat the same thing for the 'bill_depth' variable. Write a code to visualize the distribution of that variable.

In [8]:
```
plot(new_df$bill_depth_mm)
hist(new_df$bill_depth_mm, breaks = 30)
```





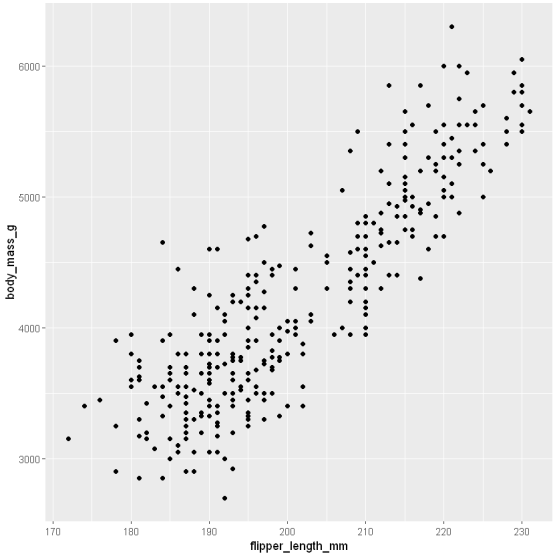
At the next step, let's consider the co-variation of some pairs of variables. 'flipper_length_mm' and 'body_mass_g' seems an interesting pair for such consideration. Write a code to visualize their joint distribution, using the 'geom_point' function or similar approaches.

How do you think 'flipper_length_mm' and 'body_mass_g' correlate with each other? Is this correlation strongly positive, positive, not eye-catching, negative, or strongly negative? Write your opinion here, before running the code.

Your Prediction of Correlation: Positive

In [11]:
```
library(ggplot2)
ggplot(new_df, aes(x=flipper_length_mm, y=body_mass_g)) + geom_point()
cor(new_df$flipper_length_mm, new_df$body_mass_g)
```

0.873210966537645



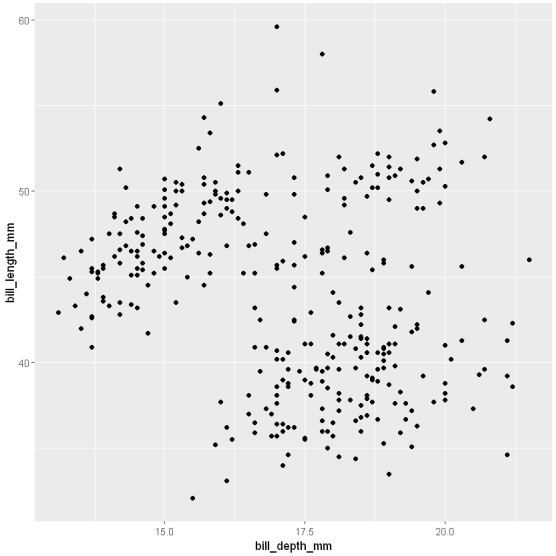
For considering another joint distribution, it is good to pick variables 'bill_length' and 'bill_depth'. Write a code for visualizing their joint distribution.

How do you think 'bill_length' and 'bill_depth' correlate with each other? Is this correlation strongly positive, positive, not eye-catching, negative, or strongly negative? Write your opinion here, before running the code.

Your Prediction of Correlation: Maybe Negative

In [12]:
```
ggplot(new_df, aes(x=bill_depth_mm, y=bill_length_mm)) + geom_point()
cor(new_df$bill_depth_mm, new_df$bill_length_mm)
```

-0.228639978056959

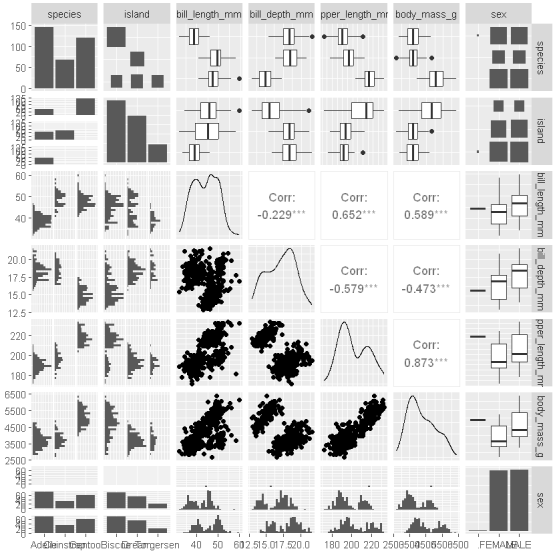


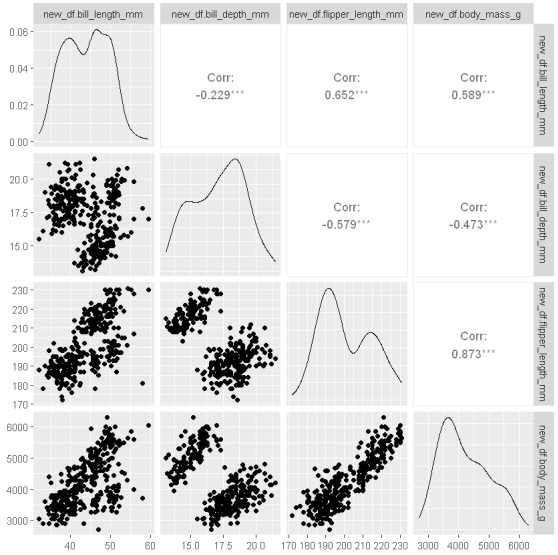
GGally: A Total Cheat!

There are even more advanced and specialized functions available in R, for helping statisticians study the distribution and co-distribution of variables more easily. So far, we have picked pairs of variables to consider based on our intuition or desires. The function 'ggpairs' in library 'GGally' draws a plot showing the distribution of every single variable, in addition to the co-distribution of every pair of variables. Use the function to draw such a plot over all numerical variables of df.

In [14]:
```
#install.packages("GGally")
library(GGally)
ggpairs(new_df)
ggpairs(data.frame(new_df$bill_length_mm, new_df$bill_depth_mm, new_df$flipper_length_mm, new_df$body_mass_g))
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```





Can you find any new interesting facts about data? Do you see a phenomenon that meets your expectations? Are there any anomalies seen in the plot? Explain your opinion.

Your Explanation
1.Correlation between body mass and flipper length is strongly positive as I predicted
2.Correlation between bill length and bill depth is positive as I predicted
3.According to graphs of bill depth per body mass, there is some kind of duality between species which shows their differences
4.According to graphs of bill depth per flipper length, there is some kind of duality between species which shows their differences
5.Both flipper length and body mass have a positive relationship among them
6.They are directly proportional to each other
7.Gentoos are the largest species

Feedback Station

Now and after initial exposure to data, it is good to summarize our observations and get 'feedback' from data.

Some properties of the data were as we expected, while others clearly violated our expectations. In particular, the non-normal distribution of 'bill_length' and 'bill_depth' cannot be ignored easily. We should consider this as an unusual phenomenon and try to continue our investigations about it.

The same is true about the correlation between bill length and depth. The usual expectation is a relatively strong positive correlation between these two, but the actual result showed little correlation between them!

However, the correlation between flipper length and body mass was as one usually expects. They showed a strong positive correlation. Can you find a proper biological-evolutional explanation for that?

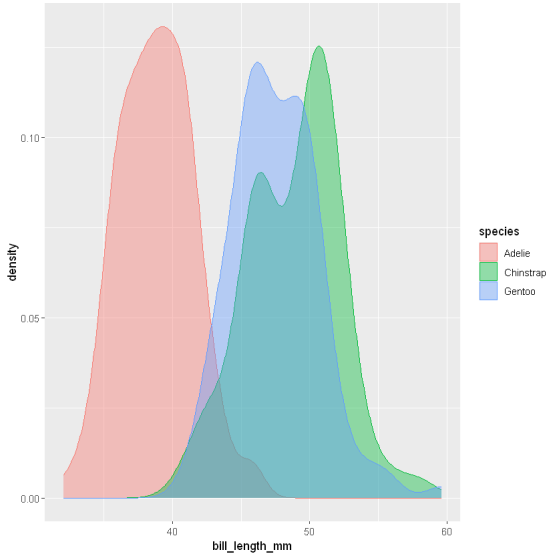
If flipper_legnth_mm >= 207, it is a Gentoo penguin (95% right)
If flipper_length_mm < 207 and bill_length_mm < 43, it is a Adelie penguin (97% right)
If flipper_length_mm < 207 and bill_length_mm >= 43, it is a Chinstrap penguin (92% right)

We continue our considerations. How can we explain the abnormal result about the correlation between bill length and depth? We normally think that if the bill length of penguin B is, say, 1.5 times the bill length of penguin A, its bill depth should also be about 1.5 times the bill depth of penguin A. This intuition implicitly includes the assumption that the 'shape' of the penguin is preserved. But this is not true, as we have different species with different 'shapes'! So what if we reconsider the correlation between bill length and depth for each specie separately? It can be a good option for further studies.

Continuing the Investigation

Write a code to draw a plot that visualizes the distribution of bill length for each specie, using the 'geom_density' function. Hint: It is good to make each 'distribution' semi-transparent for a better result.

In [15]:
```
ggplot(new_df, aes(bill_length_mm, fill=species, colour=species)) + geom_density(alpha = 0.4)
```



As we can see, the resulting distributions are relatively similar to the normal distribution. However, there are still discrepancies. We may want to end our investigation and assume these differences are caused by the error of sampling or recording data or similar reasons. But remember that this is totally in contrast with the scientific methodology. A scientist should always try to understand the truth as it really is, and **NOT** try to show that the truth is what they want it to be!

Having this said, let's find out why this difference from normal distribution still exists. To do so, write a code that selects all data points belonging to the Chinstrap specie. Then split these data points into two groups, based on which 'summit' their bill length lies. Hint: you can do the split job by comparing the bill length of each penguin with an appropriate value lying between the two 'summits'. A value of approximately 48 works fine.

In [16]:
```
funct1 <- function(x) {
    if (x > 48) {
        return (1)
    } else {
        return (0)
    }
}
funct2 <- function(x) {
    if (x > 48) {
        return ('A')
    } else {
        return ('B')
    }
}
chinstrap <- new_df[new_df$species == 'Chinstrap',]
# seperating our data to two groups by their bill_length
chinstrap$type <- sapply(chinstrap$bill_length_mm, funct1)
chinstrap$type2 <- sapply(chinstrap$bill_length_mm, funct2)
head(chinstrap, 15)
```

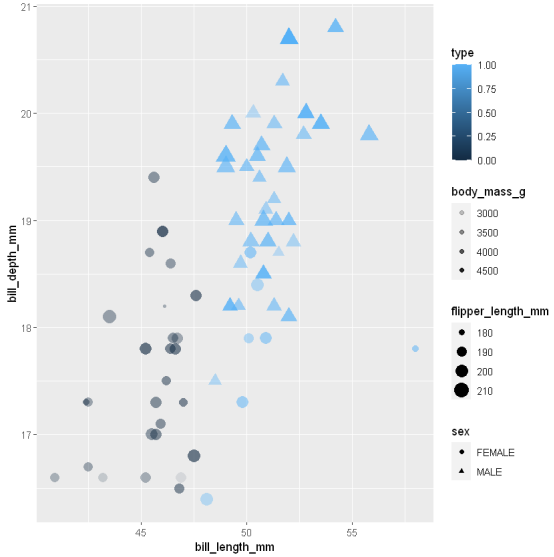
A data.frame: 15 × 9

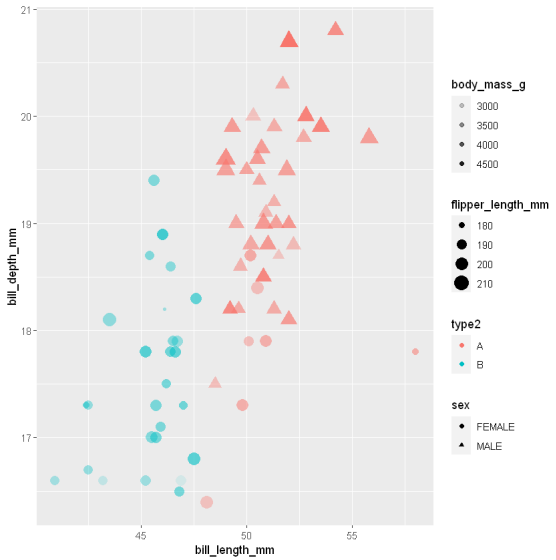
	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	type	type2
	<chr>	<chr>	<dbl>	<dbl>	<int>	<int>	<chr>	<dbl>	<chr>
153	Chinstrap	Dream	46.5	17.9	192	3500	FEMALE	0	B
154	Chinstrap	Dream	50.0	19.5	196	3900	MALE	1	A
155	Chinstrap	Dream	51.3	19.2	193	3650	MALE	1	A
156	Chinstrap	Dream	45.4	18.7	188	3525	FEMALE	0	B
157	Chinstrap	Dream	52.7	19.8	197	3725	MALE	1	A
158	Chinstrap	Dream	45.2	17.8	198	3950	FEMALE	0	B
159	Chinstrap	Dream	46.1	18.2	178	3250	FEMALE	0	B
160	Chinstrap	Dream	51.3	18.2	197	3750	MALE	1	A
161	Chinstrap	Dream	46.0	18.9	195	4150	FEMALE	0	B
162	Chinstrap	Dream	51.3	19.9	198	3700	MALE	1	A
163	Chinstrap	Dream	46.6	17.8	193	3800	FEMALE	0	B
164	Chinstrap	Dream	51.7	20.3	194	3775	MALE	1	A
165	Chinstrap	Dream	47.0	17.3	185	3700	FEMALE	0	B
166	Chinstrap	Dream	52.0	18.1	201	4050	MALE	1	A
167	Chinstrap	Dream	45.9	17.1	190	3575	FEMALE	0	B

Now it is a good idea to draw a 'universal' plot to compare the two groups. Draw a plot over Chinstrap penguins that uses the following aesthetic mapping:

- Color: The group in which the penguin lies (based on its bill length)
- Shape: Sex
- x-coordinate: Bill length
- y-coordinate: Bill depth
- Size: Flipper length
- Transparency: Body mass

In [17]:
```
# plotting with numeric types
ggplot(chinstrap,
       aes(y=bill_depth_mm,
           x=bill_length_mm,
           colour=type,
           shape=sex,
           size=flipper_length_mm,
           alpha=body_mass_g)) +
  geom_point()
# plotting with characteristics types
ggplot(chinstrap,
       aes(y=bill_depth_mm,
           x=bill_length_mm,
           colour=type2,
           shape=sex,
           size=flipper_length_mm,
           alpha=body_mass_g)) +
  geom_point()
```





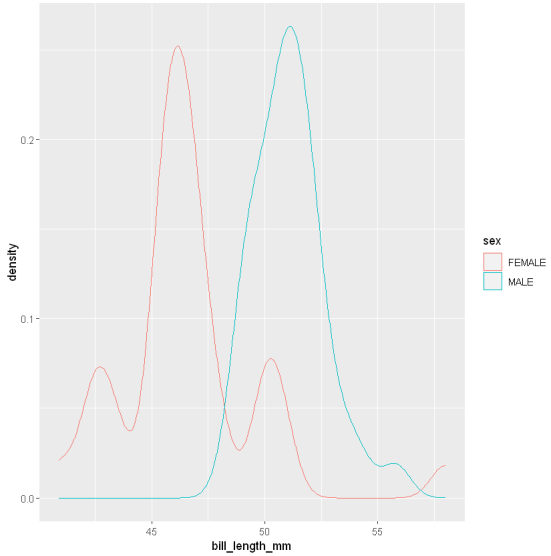
Now we can find a specific property in the above list that separates the two groups. If you did not find it, take a deeper look at the plot. After some careful observation, it will emerge. You can run the cell below to confirm your guess. We call this property Q.

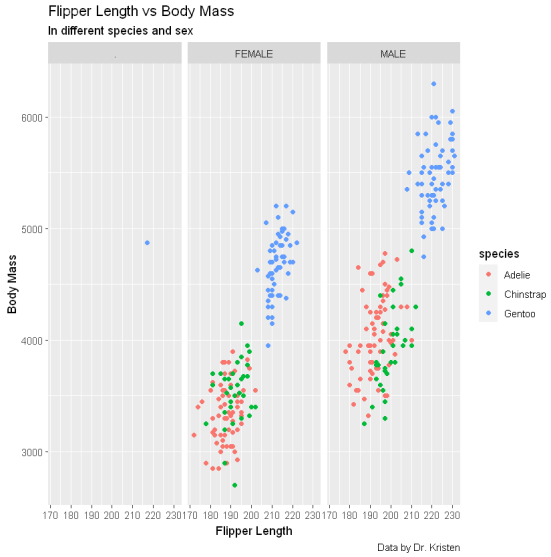
In [18]:
```
print(sapply(lapply(strsplit("selamef fo stsisnoc rehto eht dna selam fo stsisnoc ylniam spuorg eht fo eno .xes ni reffid yllareneg spuorg owt eht", NULL), rev), paste, collap
```

```
[1] "the two groups generally differ in sex. one of the groups mainly consists of males and the other consists of females"
```

Now draw a plot for chinstrap penguins, split by Q, that shows the distribution of bill length of each group. Hint: You can simply add (color = Q) in the aesthetic mapping.

In [26]:
```
ggplot(chinstrap, aes(x=bill_length_mm, color=sex)) +
    geom_density()
# extra work:
ggplot(data = new_df) +
  geom_point(mapping = aes(x= flipper_length_mm, y= body_mass_g, color= species))+
  labs(title = "Flipper Length vs Body Mass", subtitle = "In different species and sex",
       caption = "Data by Dr. Kristen", x= "Flipper Length", y="Body Mass") +
  facet_wrap(~sex)
```





You can see that this time the distributions are really closer to normal distribution. But there are still some discrepancies! Here, we terminate our investigation in this part to avoid lengthening the exercise. We have, however, gained some good understanding of the relation between bill length, bill depth, specie, and sex of penguins.

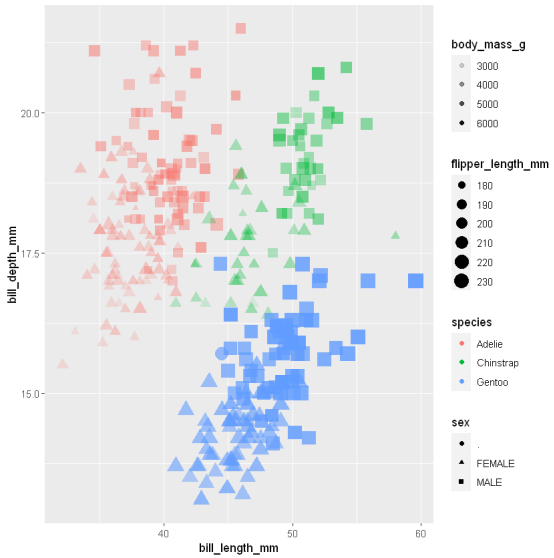
Back to the Main Course

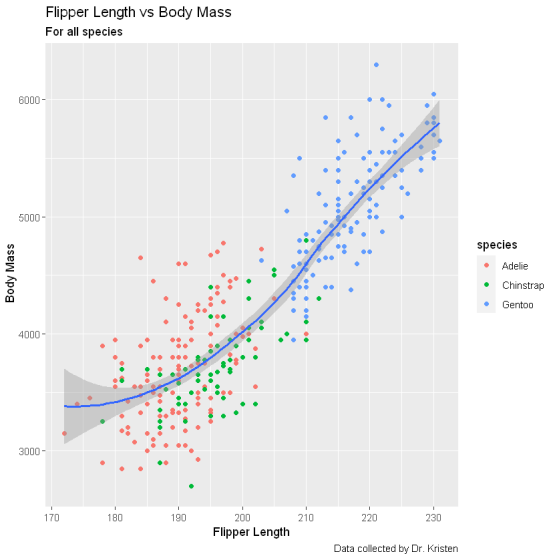
Now that we have investigated the effect of the sex parameter on the distribution of other variables for one specie, we can apply our method to the rest of the data. Using the recent plots that you have drawn, write a code to visualize all of the data in a single graph. You can use the comprehensive plot that you drew for a single specie in the previous part. One possible aesthetic mapping is like this:

- Color: Specie
- Shape: Sex
- x-Coordination: Bill length
- y-Coordination: Bill depth
- Size: Flipper length
- Transparency: Body mass

In [20]:
```
ggplot(new_df,
       aes(colour=species,
           shape=sex,
           x=bill_length_mm,
           y=bill_depth_mm,
           size=flipper_length_mm,
           alpha=body_mass_g)) +
  geom_point()
# extra work:
ggplot(data = new_df) +
  geom_point(mapping = aes(x= flipper_length_mm, y= body_mass_g, color= species )) +
  geom_smooth(mapping = aes(x= flipper_length_mm, y= body_mass_g))+
  labs(title = "Flipper Length vs Body Mass", subtitle = "For all species",
       caption = "Data collected by Dr. Kristen", x= "Flipper Length", y=" Body Mass", fontface= "Bold")
```

```
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



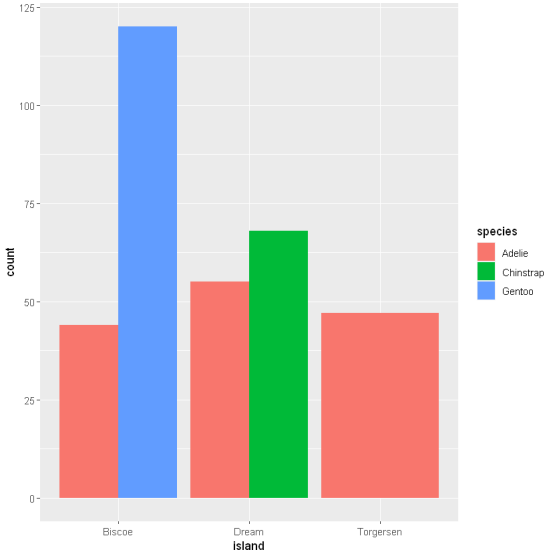


Using this new plot, we can compare the three species and observe the differences between them for every variable, and extract new patterns. Analyze the plot and describe what you observed.

Both flipper length and body mass have a positive relationship among them They are directly proportional to each other Gentoos are the largest species Males are larger than females sex which are NA are insignificant in Chinstrap species

So far, we have drawn many plots in order to extract new information and hidden facts from the dataset. But we did not consider the island variable in our exploration. Now we want to add it to our consideration. Write a code to draw a bar plot that shows the population of the three species on every island.

In [21]:
```
ggplot(new_df, aes(fill=species, x=island)) +
        geom_bar(position="dodge")
```



How is the distribution of penguins on islands? Does it make sense? Can you find a rational explanation for species having different populations on islands?

- Adelie find in all islands
- We can find Gentoo penguins just in Biscoe!
- Chinstrap are in dream

Feedback Station, Again!

We found that the non-normal distribution of bill length and depth could be partially explained as an impact of sex and specie. We drew a plot based on all data points of df to compare the basic properties of different species. In addition, we tried to explain why the population of penguins is different from island to island, although this remains a guess and further studies are needed.

Now it is possible to claim that we have obtained initial knowledge and 'insight' about the data. If this exercise was real statistical research, we were ready to start our main job. Here, to evaluate your understanding of data, make it a bit fun, and extract some 'application' from this gained knowledge, we have added a surprising section to this question: Regression Analysis!

Surprising Section: Regression Analysis!

In this part, your task is to reload df from the .csv file, put its rows containing NA values in another data frame, say df_na_rows, and try to fill those NA values based on what you understand from the whole data frame. For example, if a datapoint has the 'sex' parameter missed, you can try to retrieve that piece of information by considering the bill length, bill depth, and specie of that penguin.

In [59]:
```
df_na_rows <- df[rowSums(is.na(df)) > 0,]
df_na_rows
library('dplyr')
adelie_df <- filter(df_na_rows, species == "Adelie")
gentoo_df <- filter(df_na_rows, species == "Gentoo")
adelie_df$sex <- with(adelie_df, ifelse(bill_length_mm <= 38, 'Female','Male'))
gentoo_df$sex <- with(gentoo_df, ifelse(bill_length_mm <= 47, 'Female','Male'))
adelie_df
gentoo_df
newdf <- rbind(adelie_df, gentoo_df)
newdf
df_na_rows <- newdf
```

A data.frame: 10 × 7

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
	<chr>	<chr>	<dbl>	<dbl>	<int>	<int>	<chr>
4	Adelie	Torgersen	NA	NA	NA	NA	NA
9	Adelie	Torgersen	34.1	18.1	193	3475	NA
10	Adelie	Torgersen	42.0	20.2	190	4250	NA
11	Adelie	Torgersen	37.8	17.1	186	3300	NA
12	Adelie	Torgersen	37.8	17.3	180	3700	NA
48	Adelie	Dream	37.5	18.9	179	2975	NA
247	Gentoo	Biscoe	44.5	14.3	216	4100	NA
287	Gentoo	Biscoe	46.2	14.4	214	4650	NA
325	Gentoo	Biscoe	47.3	13.8	216	4725	NA
340	Gentoo	Biscoe	NA	NA	NA	NA	NA

A data.frame: 6 × 7

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
<chr>	<chr>	<dbl>	<dbl>	<int>	<int>	<chr>
Adelie	Torgersen	NA	NA	NA	NA	NA
Adelie	Torgersen	34.1	18.1	193	3475	Female
Adelie	Torgersen	42.0	20.2	190	4250	Male
Adelie	Torgersen	37.8	17.1	186	3300	Female
Adelie	Torgersen	37.8	17.3	180	3700	Female
Adelie	Dream	37.5	18.9	179	2975	Female

A data.frame: 4 × 7

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
<chr>	<chr>	<dbl>	<dbl>	<int>	<int>	<chr>
Gentoo	Biscoe	44.5	14.3	216	4100	Female
Gentoo	Biscoe	46.2	14.4	214	4650	Female
Gentoo	Biscoe	47.3	13.8	216	4725	Male
Gentoo	Biscoe	NA	NA	NA	NA	NA

A data.frame: 10 × 7

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
<chr>	<chr>	<dbl>	<dbl>	<int>	<int>	<chr>
Adelie	Torgersen	NA	NA	NA	NA	NA
Adelie	Torgersen	34.1	18.1	193	3475	Female
Adelie	Torgersen	42.0	20.2	190	4250	Male
Adelie	Torgersen	37.8	17.1	186	3300	Female
Adelie	Torgersen	37.8	17.3	180	3700	Female
Adelie	Dream	37.5	18.9	179	2975	Female
Gentoo	Biscoe	44.5	14.3	216	4100	Female
Gentoo	Biscoe	46.2	14.4	214	4650	Female
Gentoo	Biscoe	47.3	13.8	216	4725	Male
Gentoo	Biscoe	NA	NA	NA	NA	NA

In [60]:
```
df1 <- new_df[new_df$species == "Adelie" & new_df$island == "Torgersen", ]
df2 <- new_df[new_df$species == "Gentoo" & new_df$island == "Biscoe", ]
df_na_rows$bill_length_mm[1] <- mean(df1$bill_length_mm)
df_na_rows$bill_depth_mm[1] <- mean(df1$bill_depth_mm)
df_na_rows$flipper_length_mm[1] <- mean(df1$flipper_length_mm)
df_na_rows$body_mass_g[1] <- mean(df1$body_mass_g)
df_na_rows$sex[1] <- ifelse(df_na_rows$bill_length_mm[1] > 38, "Male", "Female")
# ------------------------------------
df_na_rows$bill_length_mm[10] <- mean(df2$bill_length_mm)
df_na_rows$bill_depth_mm[10] <- mean(df2$bill_depth_mm)
df_na_rows$flipper_length_mm[10] <- mean(df2$flipper_length_mm)
df_na_rows$body_mass_g[10] <- mean(df2$body_mass_g)
df_na_rows$sex[10] <- ifelse(df_na_rows$bill_length_mm[10] > 47, "Male", "Female")
df_na_rows
```

A data.frame: 10 × 7

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
Adelie	Torgersen	39.0383	18.45106	191.5319	3708.511	Male
Adelie	Torgersen	34.1000	18.10000	193.0000	3475.000	Female
Adelie	Torgersen	42.0000	20.20000	190.0000	4250.000	Male
Adelie	Torgersen	37.8000	17.10000	186.0000	3300.000	Female
Adelie	Torgersen	37.8000	17.30000	180.0000	3700.000	Female
Adelie	Dream	37.5000	18.90000	179.0000	2975.000	Female
Gentoo	Biscoe	44.5000	14.30000	216.0000	4100.000	Female
Gentoo	Biscoe	46.2000	14.40000	214.0000	4650.000	Female
Gentoo	Biscoe	47.3000	13.80000	216.0000	4725.000	Male
Gentoo	Biscoe	47.5425	15.00250	217.2333	5090.625	Male

Question 2 - Exploring on Your Own

In this question, we have some interesting real-world datasets. You are to pick one of these datasets and perform an EDA using what you have learned from the previous question.

In [1]:
```
df <- read.csv("insurance.csv")
head(df, 10)
```

A data.frame: 10 × 7

	age	sex	bmi	children	smoker	region	charges
	<int>	<chr>	<dbl>	<int>	<chr>	<chr>	<dbl>
1	19	female	27.900	0	yes	southwest	16884.924
2	18	male	33.770	1	no	southeast	1725.552
3	28	male	33.000	3	no	southeast	4449.462
4	33	male	22.705	0	no	northwest	21984.471
5	32	male	28.880	0	no	northwest	3866.855
6	31	female	25.740	0	no	southeast	3756.622
7	46	female	33.440	1	no	southeast	8240.590
8	37	female	27.740	3	no	northwest	7281.506
9	37	male	29.830	2	no	northeast	6406.411
10	60	female	25.840	0	no	northwest	28923.137

In [30]:
```
str(df)
summary(df)
skim(df)
```

```
'data.frame':   1338 obs. of  7 variables:
 $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
 $ sex     : chr  "female" "male" "male" "male" ...
 $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
 $ children: int  0 1 3 0 0 0 1 3 2 0 ...
 $ smoker  : chr  "yes" "no" "no" "no" ...
 $ region  : chr  "southwest" "southeast" "southeast" "northwest" ...
 $ charges : num  16885 1726 4449 21984 3867 ...

      age            sex                 bmi           children
 Min.   :18.00   Length:1338        Min.   :15.96   Min.   :0.000
 1st Qu.:27.00   Class :character   1st Qu.:26.30   1st Qu.:0.000
 Median :39.00   Mode  :character   Median :30.40   Median :1.000
 Mean   :39.21                      Mean   :30.66   Mean   :1.095
 3rd Qu.:51.00                      3rd Qu.:34.69   3rd Qu.:2.000
 Max.   :64.00                      Max.   :53.13   Max.   :5.000
    smoker             region             charges
 Length:1338        Length:1338        Min.   : 1122
 Class :character   Class :character   1st Qu.: 4740
 Mode  :character   Mode  :character   Median : 9382
                                       Mean   :13270
                                       3rd Qu.:16640
                                       Max.   :63770

── Data Summary ────────────────────────
                           Values
Name                       df
Number of rows             1338
Number of columns          7
_________________________
Column type frequency:
  character                3
  numeric                  4
_________________________
Group variables            None

── Variable type: character ────────────────────────────────────────────────────
  skim_variable n_missing complete_rate min max empty n_unique whitespace
1 sex                   0             1   4   6     0        2          0
2 smoker                0             1   2   3     0        2          0
3 region                0             1   9   9     0        4          0

── Variable type: numeric ──────────────────────────────────────────────────────
  skim_variable n_missing complete_rate     mean       sd     p0    p25    p50
1 age                   0             1    39.2     14.0    18     27     39
2 bmi                   0             1    30.7      6.10   16.0   26.3   30.4
3 children              0             1     1.09     1.21    0      0      1
4 charges               0             1 13270.   12110.   1122.  4740.  9382.
      p75    p100 hist
1    51      64   ▇▅▅▅▇▇
2    34.7    53.1 ▁▃▇▇▃▁
3     2       5   ▇▂▁▂▁▁
4 16640.  63770.  ▇▂▁▁▁▁

Warning message in is.null(text_repr) || nchar(text_repr) == 0L:
"'length(x) = 17 > 1' in coercion to 'logical(1)'"
```

A skim_df: 7 × 17

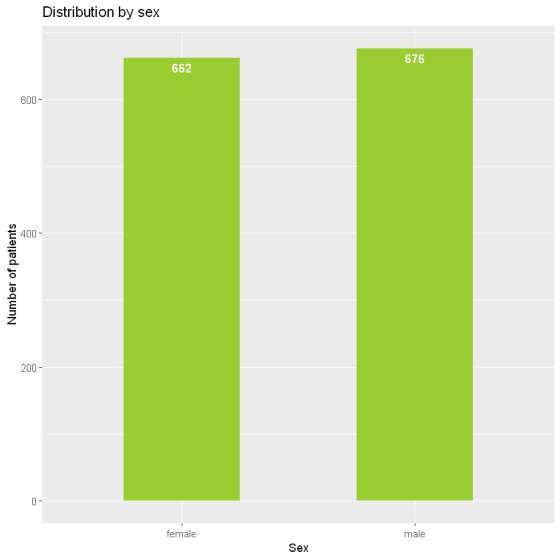
	skim_type	skim_variable	n_missing	complete_rate	character.min	character.max	character.empty	character.n_unique	character.whitespace	numeric.mean	numeric.sd	numeric.p0	nume
	<chr>	<chr>	<int>	<dbl>	<int>	<int>	<int>	<int>	<int>	<dbl>	<dbl>	<dbl>	
1	character	sex	0	1	4	6	0	2	0	NA	NA	NA	
2	character	smoker	0	1	2	3	0	2	0	NA	NA	NA	
3	character	region	0	1	9	9	0	4	0	NA	NA	NA	
4	numeric	age	0	1	NA	NA	NA	NA	NA	39.207025	14.049960	18.000	2
5	numeric	bmi	0	1	NA	NA	NA	NA	NA	30.663397	6.098187	15.960	2
6	numeric	children	0	1	NA	NA	NA	NA	NA	1.094918	1.205493	0.000	
7	numeric	charges	0	1	NA	NA	NA	NA	NA	13270.422265	12110.011237	1121.874	474

In [32]:
```
sum(is.na(df))
```

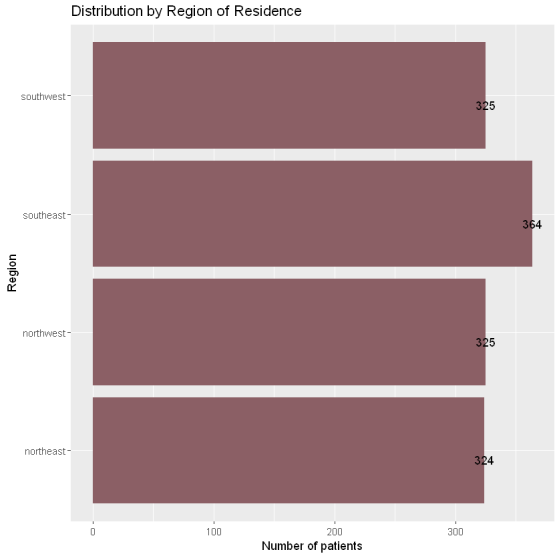
0

In [117]:
```
#It seems that there are somewhat equal number of male and female patients.
```

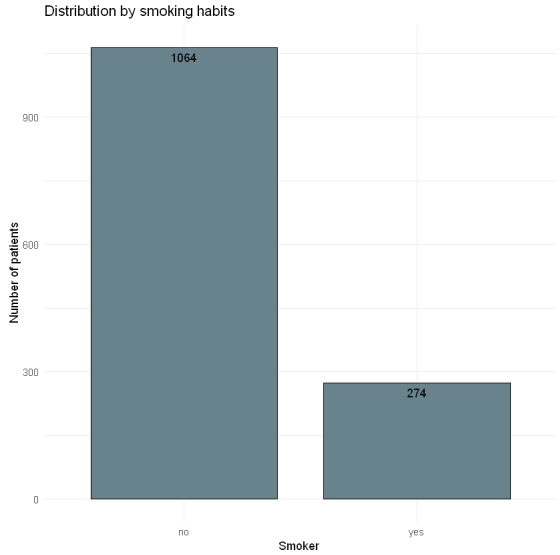
In [44]:
```
ggplot(df, aes(x = sex)) +
  geom_bar(width=0.5, fill="yellowgreen") +
  geom_text(aes(label = ..count..), stat = "count", vjust = 1.5, color = "white")+
  labs(title = "Distribution by sex",
       x="Sex", y="Number of patients")
```



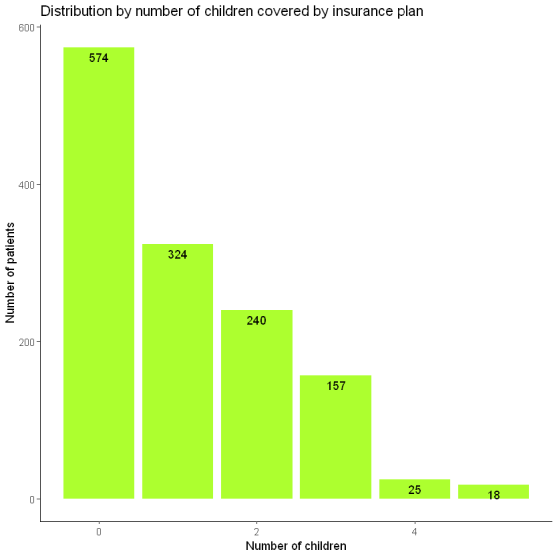
In [126]:
```
ggplot(df, aes(x=region))+
  geom_bar(fill = "lightpink4")+
  geom_text(aes(label = ..count..), stat = "count", vjust = 1.5, color = "black")+
  labs(title = "Distribution by Region of Residence",
       x="Region",y="Number of patients")+
  coord_flip()
```



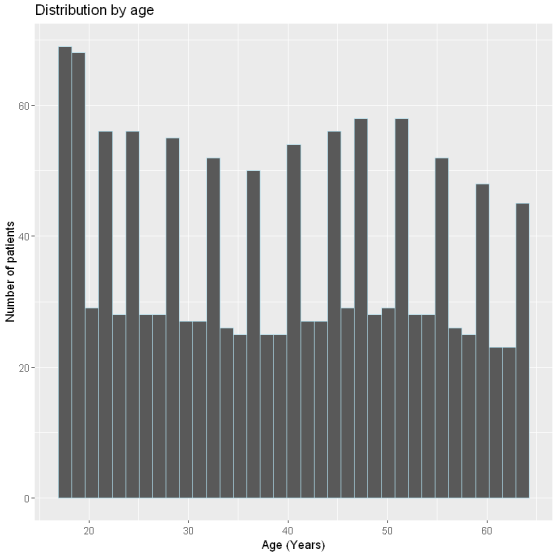
In [127]:
```
ggplot(df, aes(x=smoker))+
  geom_bar(width = 0.8, fill='lightblue4', color='black')+
  geom_text(aes(label = ..count..), stat = "count", vjust = 1.5, colour = "black")+
  labs(title = "Distribution by smoking habits",
       x="Smoker ",y="Number of patients")+
  theme_minimal()
```



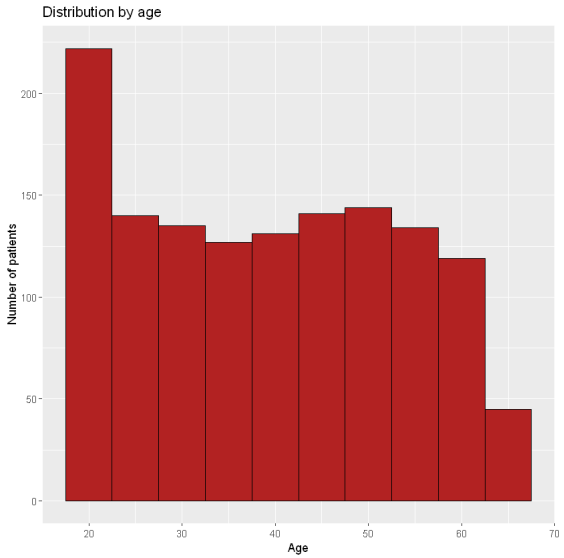
In [56]:
```
ggplot(df, aes(x=children))+
  geom_bar(fill = "greenyellow")+
  geom_text(aes(label = ..count..), stat = "count", vjust = 1.5, color = "black")+
  theme_classic()+
  labs(title = "Distribution by number of children covered by insurance plan",
       x = "Number of children", y = "Number of patients")
```



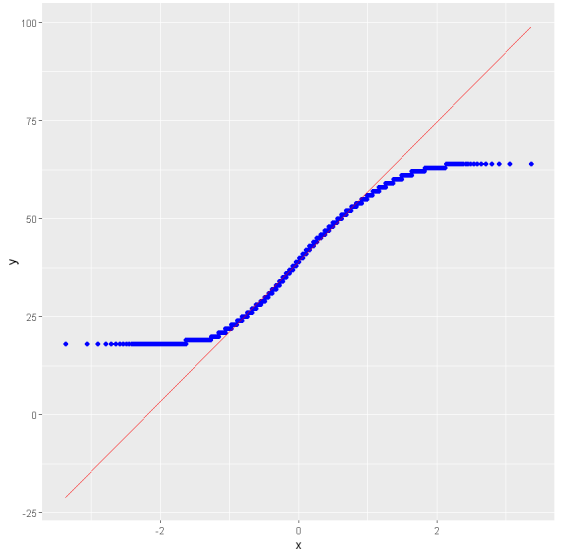
In [68]:
```
ggplot(df, aes(x=age))+
  geom_histogram(col = 'lightblue', bins = 35)+
  labs(title = "Distribution by age",
       x="Age (Years)",y="Number of patients")
```



In [70]:
```
ggplot(df, aes(x = age)) +
  geom_histogram(binwidth = 5,color = "black", fill = "firebrick")+
  labs(title = "Distribution by age",
       x="Age",y="Number of patients")
```

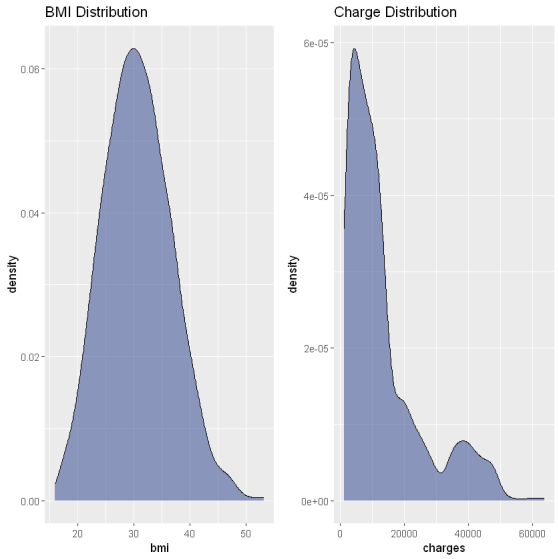


In [73]:
```
ggplot(df, aes(sample = age)) +
  geom_qq(col = 'blue') +
  geom_qq_line(col = 'red')
```



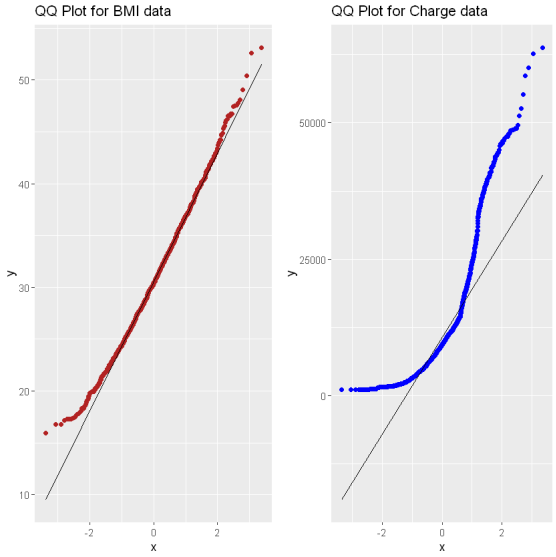
In [118]:
```
#On the x-axis are theoretical quantiles from a normal distribution while sample quantiles are along the y-axis.
#If the plotted points match up along the straight lines,
#then we say that the quantiles match and hence the data comes from a normal distribution.
#The age data does not seem to be normally distributed.
```

In [79]:
```
library(gridExtra)
p1 = ggplot(df, aes(x = bmi)) +
  geom_density(fill = "royalblue4", alpha = .5)+
  labs(title = "BMI Distribution")
p2 = ggplot(df, aes(x = charges)) +
  geom_density(fill = "royalblue4", alpha = .5)+
  labs(title = "Charge Distribution")
grid.arrange(p1, p2, ncol=2)
```



In [119]:
```
#While the bmi data seem to follow a normal distribution, charge does not. We can again check the QQ plot of these two.
```

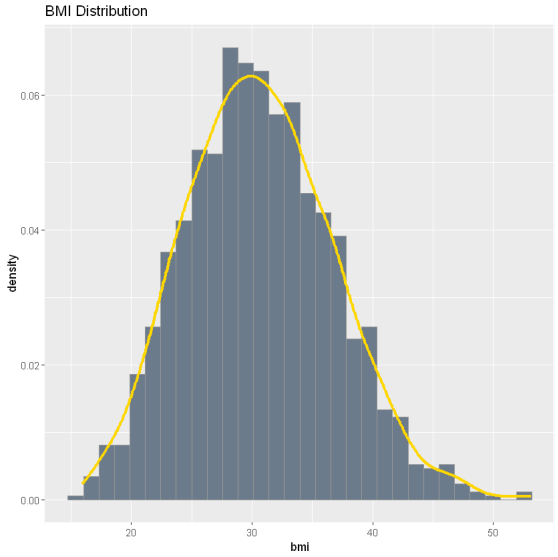
In [83]:
```
p3 = ggplot(df, aes(sample = bmi)) +
  geom_qq(col = 'firebrick') +
  geom_qq_line()+
  labs(title = "QQ Plot for BMI data")

p4 = ggplot(df, aes(sample = charges)) +
  geom_qq(col = 'blue') +
  geom_qq_line()+
  labs(title = "QQ Plot for Charge data")

grid.arrange(p3, p4, ncol=2)
```



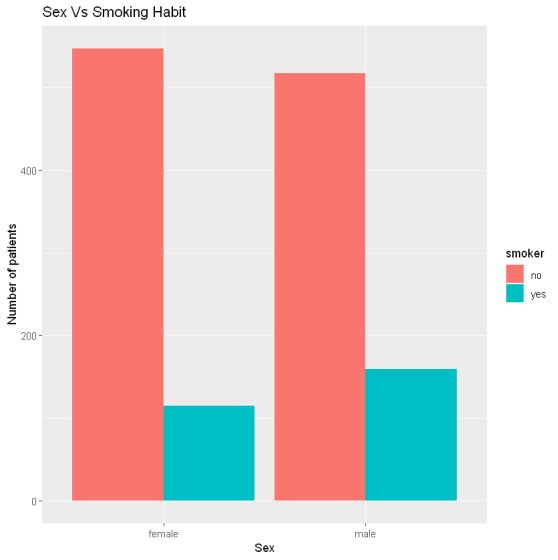
In [120]:
```
#We have further evidence for non-normality of the charge data from QQ plot.
```

In [92]:
```
ggplot(df, aes(x = bmi, y = ..density..)) +
  geom_histogram(fill = "slategray4", color = "grey60", size = .2) +
  geom_density(col = 'gold', size = 1.3)+
  labs(title = "BMI Distribution")
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

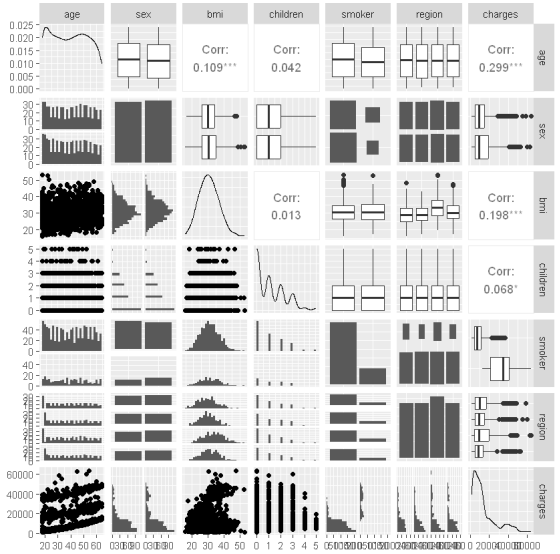


In [93]:
```
ggplot(df, aes(sex, fill = smoker))+
  geom_bar(position = "dodge")+
  labs(title = "Sex Vs Smoking Habit",
       x = "Sex", y = "Number of patients")
```



```
In [4]: #install.packages("GGally")
        library(GGally)
        ggpairs(df)
```

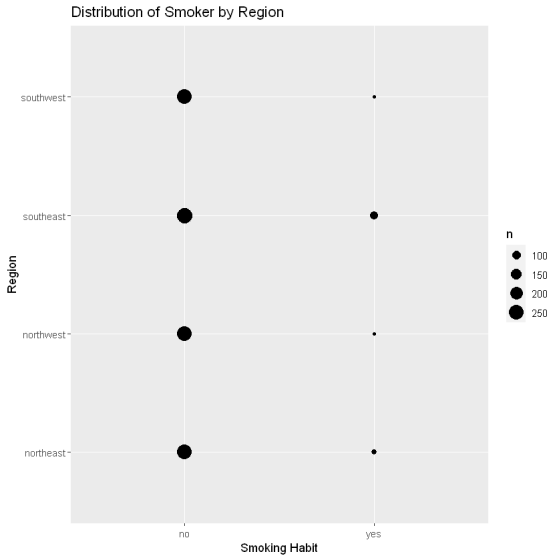
```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
In [94]: table(df$sex, df$smoker)
```

```
         no yes
  female 547 115
  male   517 159
```

In [95]:
```
ggplot(df) +
  geom_count(mapping = aes(x = smoker, y = region))+
  labs(title = "Distribution of Smoker by Region",
       x = "Smoking Habit", y = "Region")
```

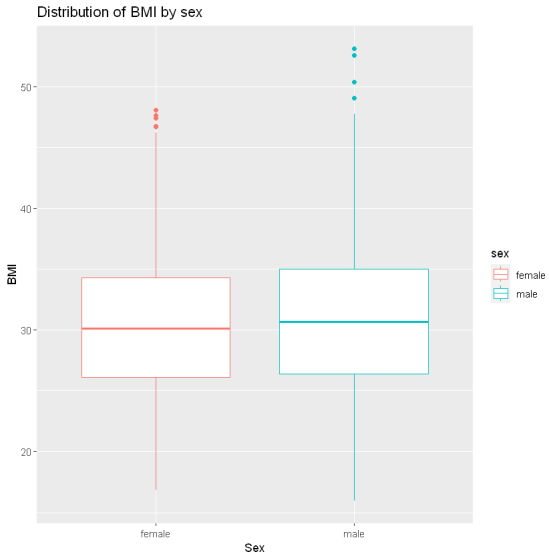


In [99]:
```
library(dplyr)
df %>%
  count(smoker, region)
```

A data.frame: 8 × 3

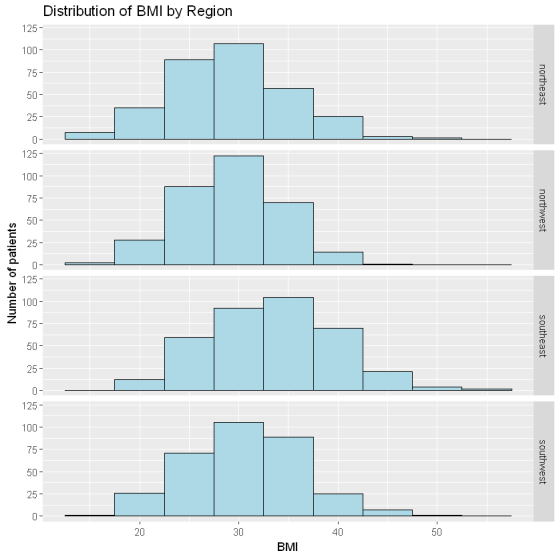
smoker	region	n
<chr>	<chr>	<int>
no	northeast	257
no	northwest	267
no	southeast	273
no	southwest	267
yes	northeast	67
yes	northwest	58
yes	southeast	91
yes	southwest	58

In [100]:
```
ggplot(df, aes(x=sex,y=bmi,color=sex))+
  geom_boxplot()+
  labs(title = "Distribution of BMI by sex",
       x = "Sex", y = "BMI")
```



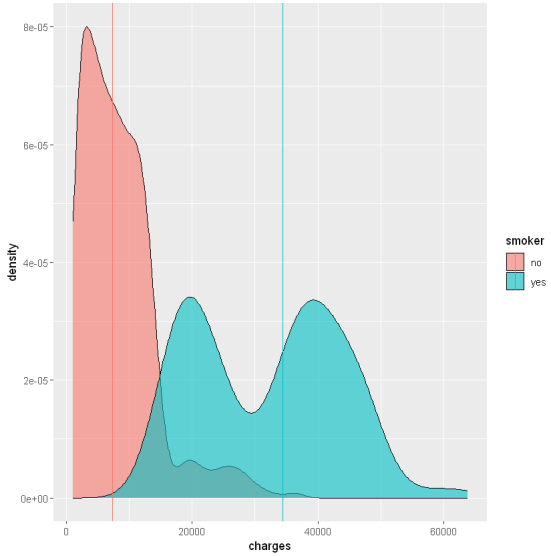
In [121]:
```
#Median BMI value for male looks slightly higher than for female
```

In [101]:
```
ggplot(df, aes(x = bmi)) +
  geom_histogram(binwidth = 5, fill = "lightblue", colour = "black") +
  labs(title = "Distribution of BMI by Region",
       x = "BMI", y = "Number of patients")+
  facet_grid(region ~ .)
```

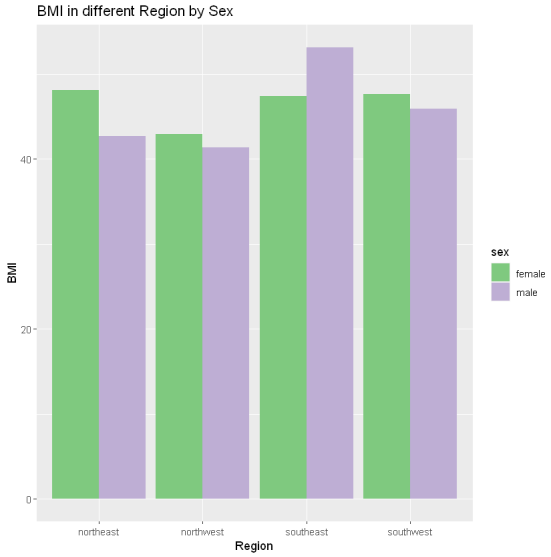


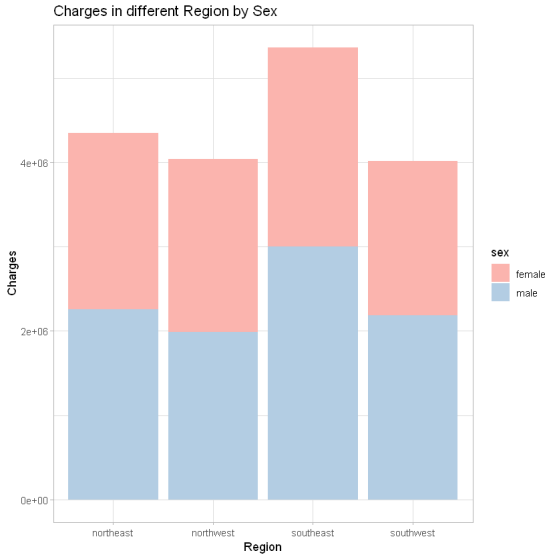
In [122]:
```
#BMI seems to be somewhat homogeneously distributed in each region.
```

In [103]:
```
med_charges <- df %>%
  group_by(smoker) %>%
  summarize(Median=median(charges))

ggplot(df, aes(charges))+
  geom_density(aes(fill=smoker),alpha=0.6)+
  geom_vline(data = med_charges, aes(xintercept = Median,color=smoker))
```



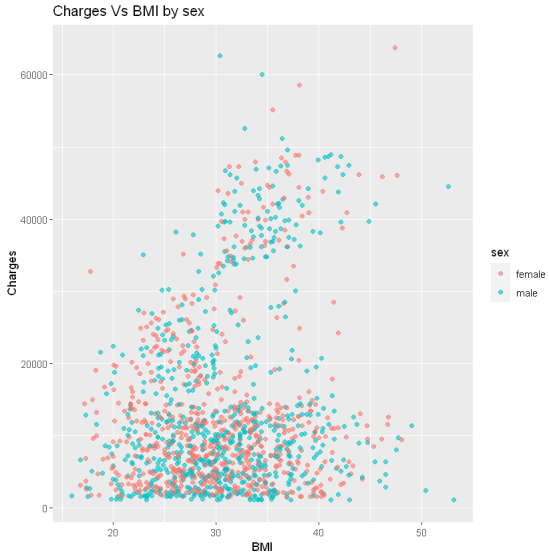
In [105]:
```
ggplot(df, aes(x = region, y = bmi, fill = sex)) +
  geom_col(position = "dodge")+
  scale_fill_brewer(palette = "Accent")+
  labs(title = "BMI in different Region by Sex",
       x = "Region", y = "BMI")

ggplot(df, aes(x = region, y = charges, fill = sex)) +
  geom_col() +
  scale_fill_brewer(palette = "Pastel1")+
  labs(title = "Charges in different Region by Sex",
       x = "Region", y = "Charges")+
  theme_light()
```





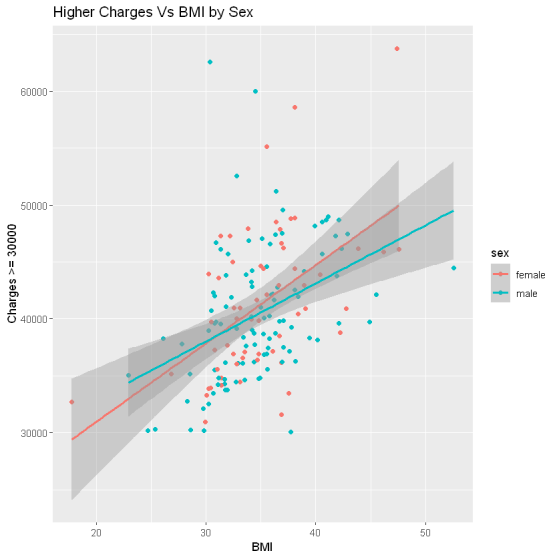
In [124]:
```
#The highest charges for males are in the southeast region while the lowest are in the northwest.
#Again, we can get a summary table of these values for all combinations.
```

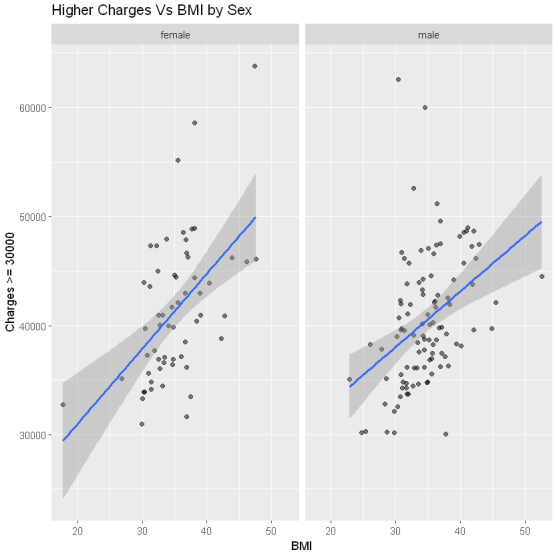
In [106]:
```
ggplot(df,aes(x=bmi,y=charges,color=sex))+
  geom_point(alpha=0.6)+
  labs(title = "Charges Vs BMI by sex",
       x = "BMI", y = "Charges")
```



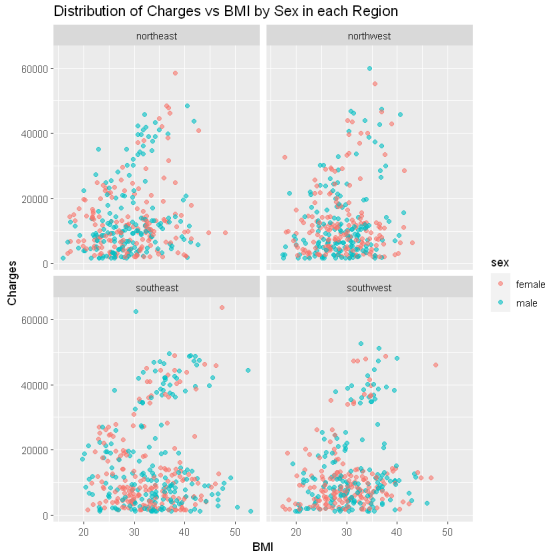
In [109]:
```
higher.charges <- df %>%
  filter(charges >= 30000)
ggplot(higher.charges,aes(x=bmi,y=charges,color=sex))+
  geom_point()+
  stat_smooth(method=lm)+
  labs(title = "Higher Charges Vs BMI by Sex",
       x = "BMI", y = "Charges >= 30000")
ggplot(higher.charges,aes(x=bmi,y=charges))+
  geom_point(alpha=0.5)+
  stat_smooth(method = lm)+
  facet_grid(.~sex)+
  labs(title = "Higher Charges Vs BMI by Sex",
       x = "BMI",
       y = "Charges >= 30000")
```

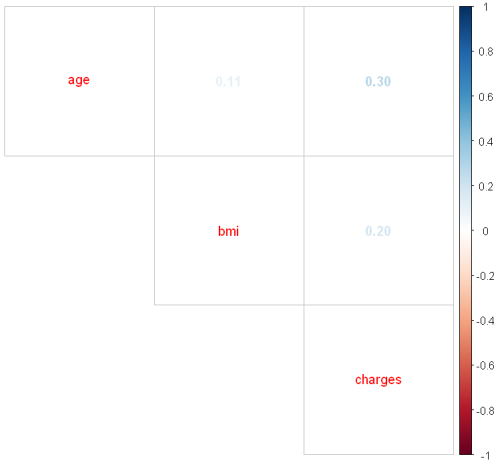
```
`geom_smooth()` using formula = 'y ~ x'
`geom_smooth()` using formula = 'y ~ x'
```



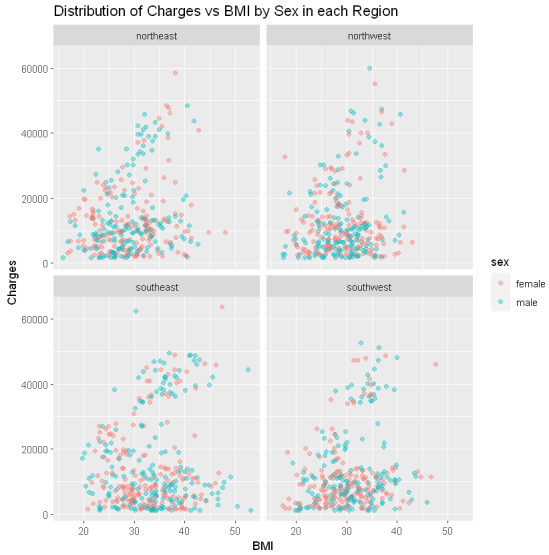

Higher Charges Vs BMI by Sex
female
male
Charges >= 30000
60000
50000
40000
30000
20
30
40
50
BMI

In [113]:
```
ggplot(df,aes(x=bmi,y=charges,color=sex))+
  geom_point(alpha = 0.6)+
  facet_wrap(~region)+
  labs(title = "Distribution of Charges vs BMI by Sex in each Region",
       x = "BMI",
       y = "Charges")
# correlation:
library(corrplot)
corr<-cor(df[,c(1,3,7)])
corrplot(corr, type="upper", method="number", tl.pos="d")
```





In [125]:
```
ggplot(df, aes(x=bmi,y=charges,color=sex))+
  geom_point(alpha = 0.4)+
  facet_wrap(~region)+
  labs(title = "Distribution of Charges vs BMI by Sex in each Region",
       x = "BMI",
       y = "Charges")
```



In [114]:
```
#The distribution looks similar in each region
```

Datasets

Dataset 1: Forbes Highest Paid Athletes 1990-2020

Link to the dataset: https://www.kaggle.com/datasets/parulpandey/forbes-highest-paid-athletes-19902019 (https://www.kaggle.com/datasets/parulpandey/forbes-highest-paid-athletes-19902019)

*The data has been extracted from topendsports.com website

Dataset 2: IT Salary Survey for EU region(2018-2020)

Link to the dataset: https://www.kaggle.com/datasets/parulpandey/2020-it-salary-survey-for-eu-region (https://www.kaggle.com/datasets/parulpandey/2020-it-salary-survey-for-eu-region)

Dataset 3: Nutrition facts for Starbucks Menu

Link to the dataset: https://www.kaggle.com/datasets/starbucks/starbucks-menu (https://www.kaggle.com/datasets/starbucks/starbucks-menu)

*Food composition data is in the public domain, but product names marked with ® or ™ remain the registered trademarks of Starbucks.

Dataset 4: Harry Potter Dataset

Link to the dataset: https://www.kaggle.com/datasets/gulsahdemiryurek/harry-potter-dataset?select=Characters.csv (https://www.kaggle.com/datasets/gulsahdemiryurek/harry-potter-dataset?select=Characters.csv)

*The other data were collected from pottermore.com and https://harrypotter.fandom.com/wiki/Main_Page (https://harrypotter.fandom.com/wiki/Main_Page)

Dataset 5: Netflix Movies and TV Shows

Link to the dataset: https://www.kaggle.com/datasets/shivamb/netflix-shows (https://www.kaggle.com/datasets/shivamb/netflix-shows)

Dataset 6: Students Performance in Exams

Link to the dataset: https://www.kaggle.com/datasets/spscientist/students-performance-in-exams (https://www.kaggle.com/datasets/spscientist/students-performance-in-exams)

Dataset 7: Medical Cost Personal Datasets

Link to the dataset: https://www.kaggle.com/datasets/mirichoi0218/insurance (https://www.kaggle.com/datasets/mirichoi0218/insurance)

*The dataset is available on GitHub https://github.com/stedy/Machine-Learning-with-R-datasets (https://github.com/stedy/Machine-Learning-with-R-datasets)