《计算与编程》第三次作业报告

王超 12031012

第一题

思路和结果：

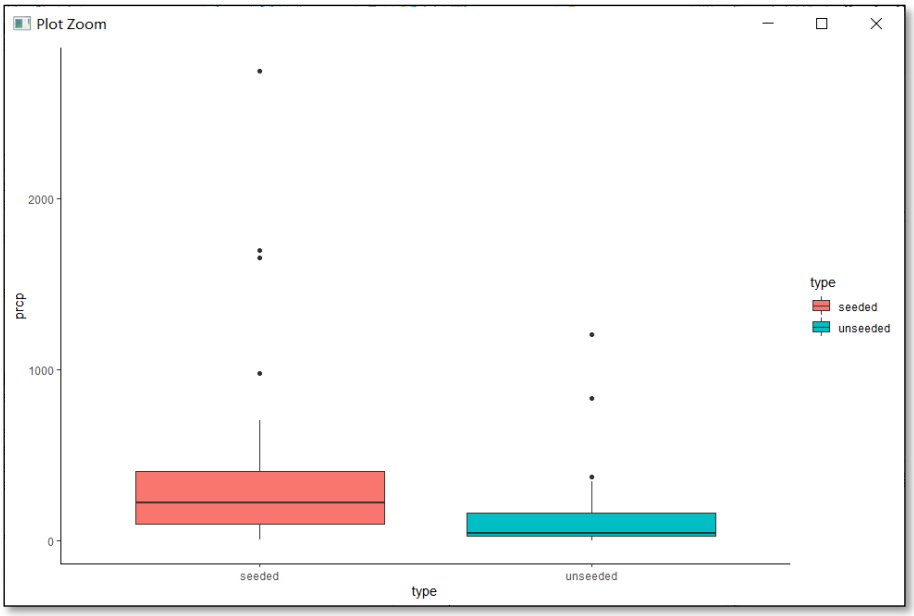
1.1

首先根据题意生成需要的数据，选择在 Excel 中生成初始数据，经过 R 处理后数据如下图所示：

```
> data1
       type   prcp
1  unseeded 1202.6
2  unseeded  830.1
3  unseeded  372.4
4  unseeded  345.5
5  unseeded  321.2
6  unseeded  244.3
```

```
`summarise()` ungrouping output (override with `.groups` argument)
# A tibble: 2 x 4
  type      count mean_prcp sd_prcp
  <chr>     <int>     <dbl>   <dbl>
1 seeded       26      442.    651.
2 unseeded     26      165.    278.
```

其中 Type 为当天是否向云层中注入碘化银，共有两类，seeded 和 unseeded，此外 prcp 为该天的降雨量。再用 boxplot 函数做出不同类型下的降雨量的箱型图，结果如下图所示：



从定性的角度来说，注入碘化银似乎会对降雨量有影响，云层中注入碘化银的那组整体降雨量有一定的提高，降雨量的平均值也升高了，但还不能说注入碘化银可以明显的提高降雨量，我们还需要进行进一步的分析。

1.2

为了探究云层注入碘化银对降雨量的影响，首先分别计算两组各自降雨量的均值和

方差，结果如下所示：

```
`summarise()` ungrouping output (override with `.groups` argument)
# A tibble: 2 x 4
  type      count mean_prcp sd_prcp
  <chr>     <int>     <dbl>   <dbl>
1 seeded       26      442.    651.
2 unseeded     26      165.    278.
```

向大气中注入碘化银后降雨量的均值有了很大的提高，说明向大气中注入碘化银会再一定程度上增加降雨，再对其进行单向方差分析，结果如下所示：

```
> summary(anova_one_way)
            Df   Sum Sq Mean Sq F value Pr(>F)
type         1  1000360 1000360   3.993 0.0511 .
Residuals   50 12525457  250509
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

P=0.0511，0.05 < P < 0.1 因此，我们不能在 95%的置信区间说明在云层注入碘化银会增加降雨，但是我们可以在 90%的置信区间内说明向大气中注入碘化银会影响降雨，并在一定程度上增加降雨。

第二题

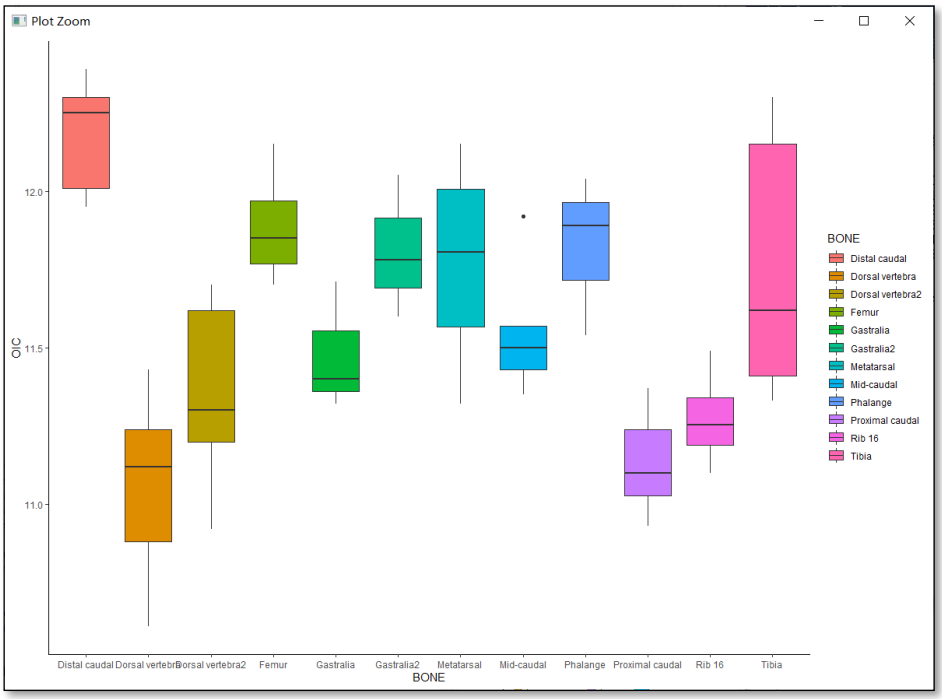
思路和结果：

首先把用 Excel 生成我们所需要的数据，包括 12 种不同部位的骨头类型以及其相应的氧的同位素组成数据，再导入到 R 中，处理后的数据如下图所示：

```
> data2
                BONE   OIC
1            Rib 16 11.10
2            Rib 16 11.22
3            Rib 16 11.29
4            Rib 16 11.49
5         Gastralia 11.32
6         Gastralia 11.40
7         Gastralia 11.71
8        Gastralia2 11.60
9        Gastralia2 11.78
10       Gastralia2 12.05
11  Dorsal vertebra 10.61
12  Dorsal vertebra 10.88
13  Dorsal vertebra 11.12
14  Dorsal vertebra 11.24
15  Dorsal vertebra 11.43
```

先做出不同部位的骨头相应的 OIC（Oxygen isotopic composition）值的箱型图，再计算不同骨头的 OIC 均值和方差，结果如下所示，结果显示不同部位的骨头的 OIC 均值是存在差异的。

```
# A tibble: 12 x 4
   BONE             count mean_OIC sd_OIC
   <chr>            <int>    <dbl>  <dbl>
 1 Distal caudal        5     12.2  0.191
 2 Dorsal vertebra      5     11.1  0.319
 3 Dorsal vertebra2     5     11.3  0.318
 4 Femur                4     11.9  0.195
 5 Gastralia            3     11.5  0.206
 6 Gastralia2           3     11.8  0.226
 7 Metatarsal           4     11.8  0.364
 8 Mid-caudal           5     11.6  0.220
 9 Phalange             3     11.8  0.257
10 Proximal caudal      6     11.1  0.166
11 Rib 16               4     11.3  0.163
12 Tibia                5     11.8  0.439
```



由于已知脊椎动物骨磷酸盐的氧同位素组成与骨骼形成时的体温有关，不同骨骼部位的均值差异表明整个身体的温度不恒定，在恒温动物中，磷酸盐的氧同位素组成预期会有较小的差值。为了说明恐龙是否为恒温动物，对其进行单因素方差分析，结果如下图所示，

```
> summary(anova_one_way)
             Df Sum Sq Mean Sq F value   Pr(>F)
BONE         11  6.067  0.5516   7.427 9.73e-07 ***
Residuals    40  2.971  0.0743
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

结果显示 P=9.73e-07 < 0.01，因此我们在 95%的置信区间，甚至是 99.9%的置信区间说明该恐龙不同部位的骨头内磷酸盐的氧同位素组成差值较大，因此说明恐龙是变温动物（冷血动物）。

第三题

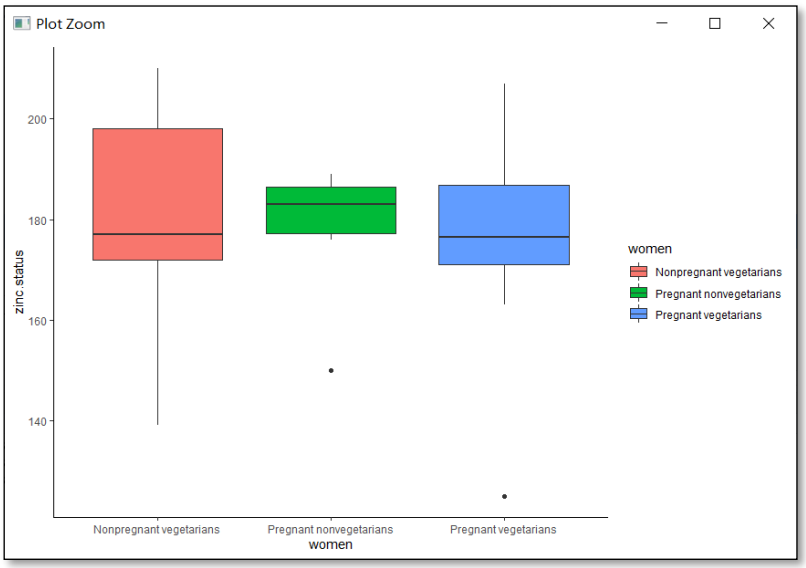
思路和结果：

首先把用 Excel 生成我们所需要的数据，再导入到 R 中，处理后的数据如下图所示，其中 women 列分为 Pregnant nonvegetarians，Pregnant vegetarians 和 Nonpregnant vegetarians 三类，pregnant 列分为 Pregnant 和 Nonpregnant 两类，vegetarians 列分为 Nonvegetarians 和 Vegetarians 两类。

```
> data3
                    women   pregnant    vegetarians zinc.status
1  Pregnant nonvegetarians   pregnant nonvegetarians         185
2  Pregnant nonvegetarians   pregnant nonvegetarians         189
3  Pregnant nonvegetarians   pregnant nonvegetarians         187
4  Pregnant nonvegetarians   pregnant nonvegetarians         181
5  Pregnant nonvegetarians   pregnant nonvegetarians         150
6  Pregnant nonvegetarians   pregnant nonvegetarians         176
7     Pregnant vegetarians   pregnant    vegetarians         171
8     Pregnant vegetarians   pregnant    vegetarians         174
9     Pregnant vegetarians   pregnant    vegetarians         202
10    Pregnant vegetarians   pregnant    vegetarians         171
11    Pregnant vegetarians   pregnant    vegetarians         207
12    Pregnant vegetarians   pregnant    vegetarians         125
13    Pregnant vegetarians   pregnant    vegetarians         189
14    Pregnant vegetarians   pregnant    vegetarians         179
15    Pregnant vegetarians   pregnant    vegetarians         163
```

再根据 women 列分别作出不同 women 类型（Pregnant nonvegetarians，Pregnant vegetarians 和 Nonpregnant vegetarians 三类）下各自 Zn 含量的均值和方差，并作出各自的箱线图，可以看出不同类型下的 Zn 含量均值差别不大。

```
`summarise()` ungrouping output (override with `.groups` argument)
# A tibble: 3 x 4
  women                  count mean_Zn sd_Zn
  <chr>                  <int>   <dbl> <dbl>
1 Nonpregnant vegetarians     5    179.  27.3
2 Pregnant nonvegetarians     6    178   14.5
3 Pregnant vegetarians       12    177.  20.9
```



由于存在是否怀孕和是否是素食主义者两类因素对 Zn 含量会有影响，因此采用双因素方差分析，如果如下图所示，图中是否怀孕和是否素食主义者两者的 P 值都远远大于 0.05，说明是否怀孕和是否为素食主义者对体内 Zn 含量没有多大的影响，因此没有证据说明怀孕的素食主义者比怀孕的非素食主义者体内的 Zn 含量更低。

```
> summary(anova_two_way)
             Df Sum Sq Mean Sq F value Pr(>F)
pregnant      1     13    12.8   0.029  0.866
vegetarians   1      3     3.4   0.008  0.931
Residuals    20   8816   440.8
```

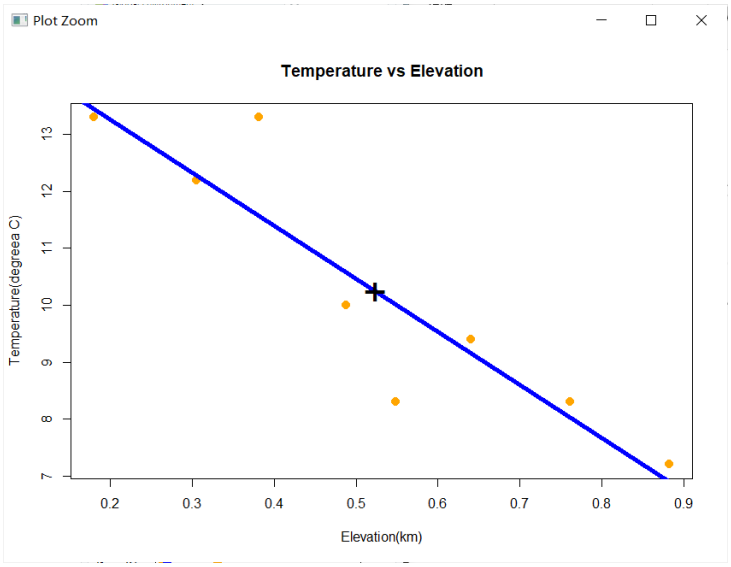
第四题

思路和结果：

首先把用 Excel 生成我们所需要的数据，再导入到 R 中，处理后的数据如下图所示，

```
> data_lm_test
      x    y
1 0.180 13.3
2 0.305 12.2
3 0.381 13.3
4 0.488 10.0
5 0.549  8.3
6 0.640  9.4
7 0.762  8.3
8 0.883  7.2
```

再用简单线性回归拟合直线，得到直线的斜率即可，最终的结果如下图所示，直线的斜率为-9.3121，说明海拔每升高 1 千米，温度降低 9.31℃，并不是降低 9.8℃。




```
> summary(fit1)

Call:
lm(formula = data_lm_test$y ~ data_lm_test$x)

Residuals:
     Min       1Q   Median       3Q      Max
-1.71254 -0.25668  0.07508  0.27763  1.72303

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     15.1249     0.9483  15.950 3.86e-06 ***
data_lm_test$x  -9.3121     1.6698  -5.577  0.00141 **
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.04 on 6 degrees of freedom
Multiple R-squared:  0.8383,    Adjusted R-squared:  0.8113
F-statistic:  31.1 on 1 and 6 DF,  p-value: 0.001411
```

第五题

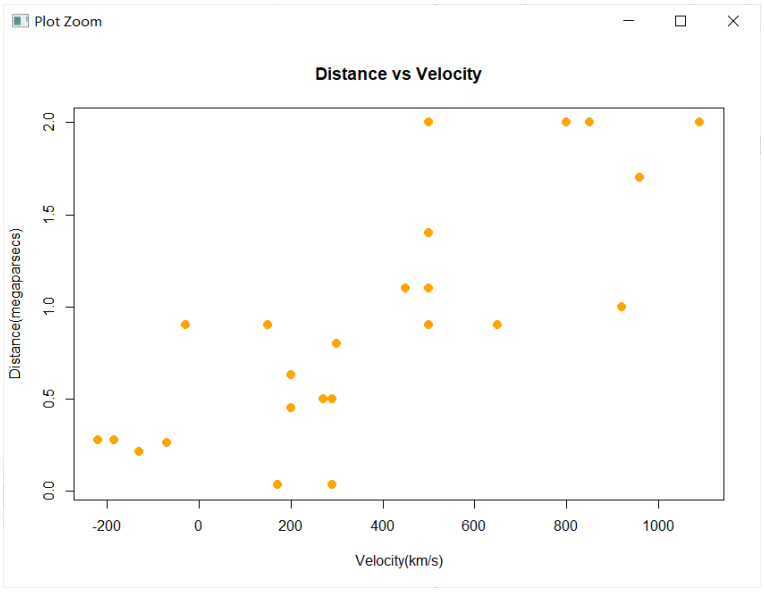
思路和结果：

5.1

首先把用 Excel 生成我们所需要的数据，再导入到 R 中，处理后的数据如下图所示，

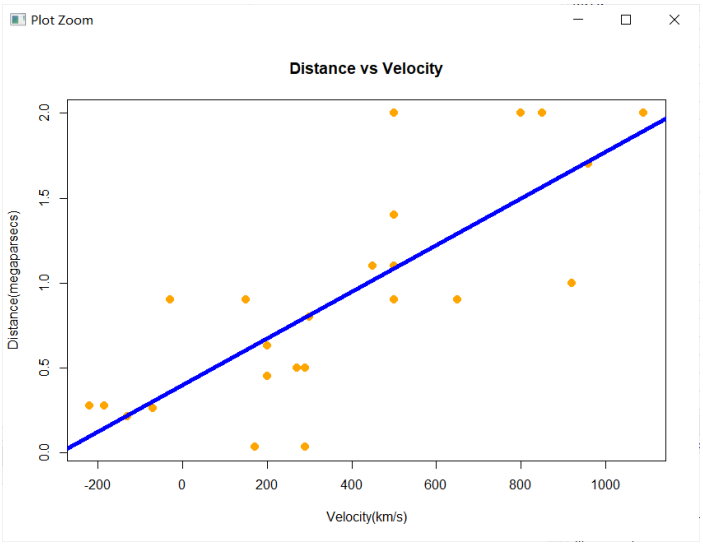
```
> data5
     Nebula Velocity Distance
1   S. Mag.      170    0.032
2   L. Mag.      290    0.034
3  NGC 6822     -130    0.214
4   NGC 598      -70    0.263
5   NGC 221     -185    0.275
6   NGC 224     -220    0.275
```

再做出散点图即可



5.2

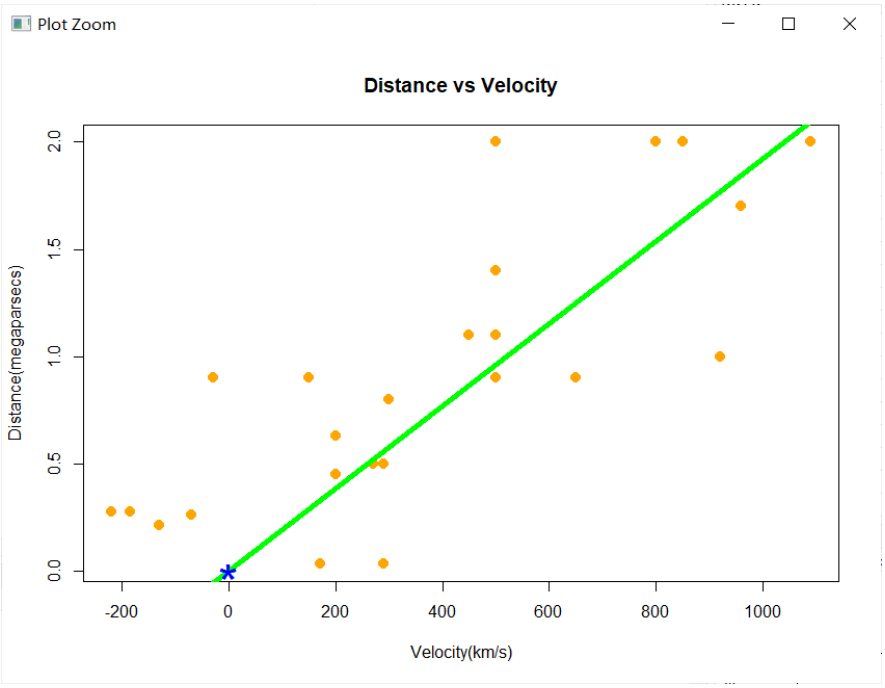
使用 lm 函数去拟合散点图，并用 abline 函数添加拟合直线



5.3

如果大爆炸理论是正确的，那么宇宙最开始是一个奇点，当宇宙爆炸开始时，所有星云都从宇宙奇点同时离开宇宙，只不过不同的星云离开奇点的速度不同，在之后的远离过程中，一直在匀速运动，因此拟合直线的截距表示的是宇宙爆炸时星云距离宇宙的距离，而此时所有星云都在奇点，因此截距必为 0；同时直线的斜率表示的是星云离开奇点往外运动的时间，这刚好是距离宇宙爆炸开始的时间，即为宇宙的年龄。通过使用以下代码来实现过原点直线的拟合，得到直线的斜率为：0.001921806，单位是（$\frac{\text{megaparsecs}}{km/s}$），换算为宇宙的年龄为：（0.001921806 * 30.9 * 10^{12} * 10^{6} ）/（365 * 24 * 60 * 60）= 18.83* 10^{8} 年 = 18.83 亿年。

```
fit3 <- lm(data5$Distance~data5$Velocity-1)
abline(fit3, lwd = 5, col = "green")
points(x=0,y=0,pch='*',cex=3 ,col='blue')
summary(fit3)
```



5.4

由于不同的星云之间的距离太远，使用传统的观测手段很难估计幸运之间的距离，通过使用改进后距离测量方法，可以减小距离的观测误差，虽然仍然有一定的误差，但是会比改进之前的误差会小，从而使得我们得到的结果也更加精确。

第六题

思路和结果：

6.1

加载数据，并对数据进行分集，随机分为 80%的训练集和 20%的测试集，再使用最佳子集回归的方法进行拟合，分别采用向前，向后和逐步子集回归方法对模型进行训练，

并查看模型的结果，最后的结果如下图所示，前进和逐步的方法模型结果都一样，即模型需要选择所有的参数。

```
sample_index <- sample(nrow(cpus),nrow(cpus)*0.80) #数据分集
cpus_train <- cpus[sample_index,]
cpus_test  <- cpus[-sample_index,]
model_1 <- lm(perf ~ syct+ mmin + mmax + cach +
                  chmin + chmax, data=cpus_train) #最佳子集回归方法
model_2=lm(perf ~ 1, data=cpus_train)
model_step_b <- step(cpus_train,direction='backward')  #backward 方法
model_step_f <- step(model_2, scope=list(lower=model_2, upper=model_1),
                     direction='forward')# forward 方法
model_step_s <- step(model_2, scope=list(lower=model_2, upper=model_1),
                     direction='both') # stepwise regression 方法
summary(model_1)
```

```
> summary(model_step_f)

Call:
lm(formula = perf ~ mmax + cach + mmin + chmax + syct, data = cpus_train)

Residuals:
    Min      1Q  Median      3Q     Max
-170.55  -24.08    5.34   23.37  429.05

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.806e+01  9.170e+00  -5.240 4.96e-07 ***
mmax         5.723e-03  7.132e-04   8.025 1.99e-13 ***
cach         6.711e-01  1.435e-01   4.678 6.11e-06 ***
mmin         1.282e-02  2.123e-03   6.038 1.04e-08 ***
chmax        1.211e+00  2.252e-01   5.377 2.63e-07 ***
syct         3.738e-02  1.829e-02   2.044   0.0426 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 58.4 on 161 degrees of freedom
Multiple R-squared:  0.8384,    Adjusted R-squared:  0.8334
F-statistic:   167 on 5 and 161 DF,  p-value: < 2.2e-16
```

```
> summary(model_step_s)

Call:
lm(formula = perf ~ mmax + cach + mmin + chmax + syct, data = cpus_train)

Residuals:
    Min      1Q  Median      3Q     Max
-170.55  -24.08    5.34   23.37  429.05

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.806e+01  9.170e+00  -5.240 4.96e-07 ***
mmax         5.723e-03  7.132e-04   8.025 1.99e-13 ***
cach         6.711e-01  1.435e-01   4.678 6.11e-06 ***
mmin         1.282e-02  2.123e-03   6.038 1.04e-08 ***
chmax        1.211e+00  2.252e-01   5.377 2.63e-07 ***
syct         3.738e-02  1.829e-02   2.044   0.0426 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 58.4 on 161 degrees of freedom
Multiple R-squared:  0.8384,    Adjusted R-squared:  0.8334
F-statistic:   167 on 5 and 161 DF,  p-value: < 2.2e-16
```

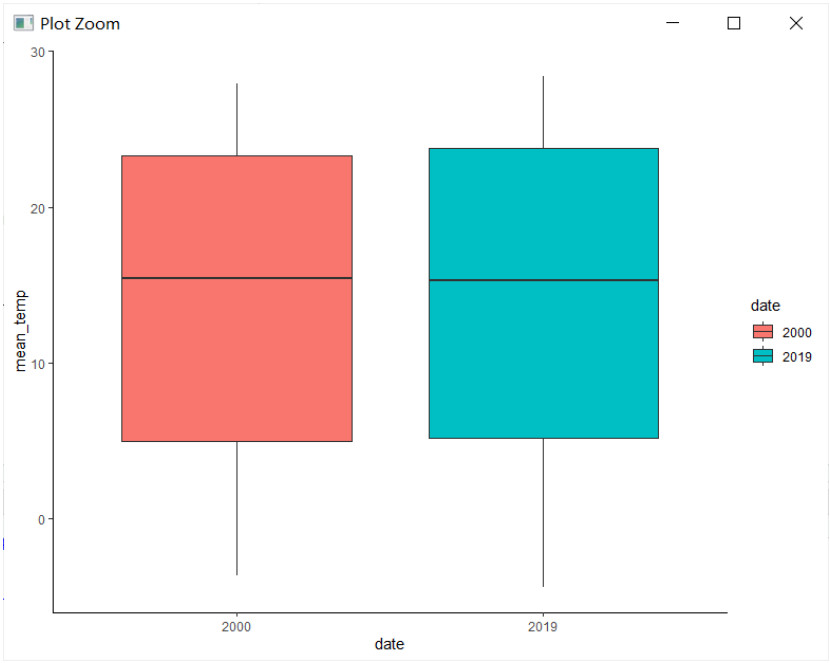
6.2

第七题

思路和结果：

7.1 探究 2000 年和 2019 年的月平均温度是否存在差异？

首先处理数据，计算得到 2000 年和 2019 年的逐月平均数据，做出不同年份月平均温度的箱线图，查看 2000 年和 2019 年各自的分布，进行 t-test 检验，查看 2000 年和 2019 年月平均温度数据差异如何。结果显示 P=0.4055＞0.05，因此认为 2000 年和 2019 年的逐月平均温度没有存在明显的统计学差异。



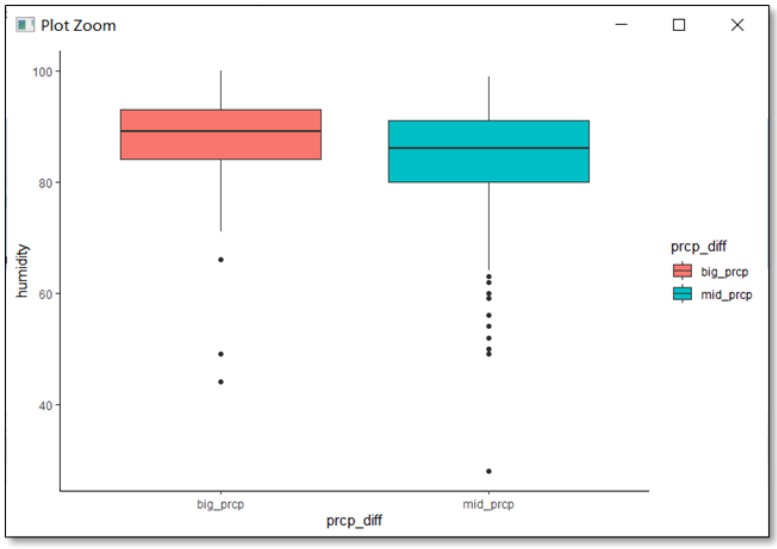

```
> t.test(box_data2000$mean_temp,box_data2019$mean_temp)

        Welch Two Sample t-test

data:  box_data2000$mean_temp and box_data2019$mean_temp
t = 0.84825, df = 21.855, p-value = 0.4055
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.259852 12.535752
sample estimates:
mean of x mean of y
 14.45517  10.81722
```

7.2 当降雨量为中雨和大雨时，对当天的空气湿度的影响如何？

我们知道降雨会当天的空气湿度，而由于无雨和小雨之间的空气湿度差异较大，因此我们想探究中雨和大雨情形下各自的空气湿度情况。首先处理数据，根据降雨量多少分别标记为：大雨和中雨，并得到对应的日均空气湿度做出不同降雨类型的箱线图，再进行单因素方差分析，探究在不同降雨量对空气湿度的影响，结果如下所示，单因素方差分析的结果中 P=0.0273 < 0.05 ，因此我们可以认为大雨和中雨会影响当天的空气湿度，同时大雨的情形下空气湿度均值要高于中雨。

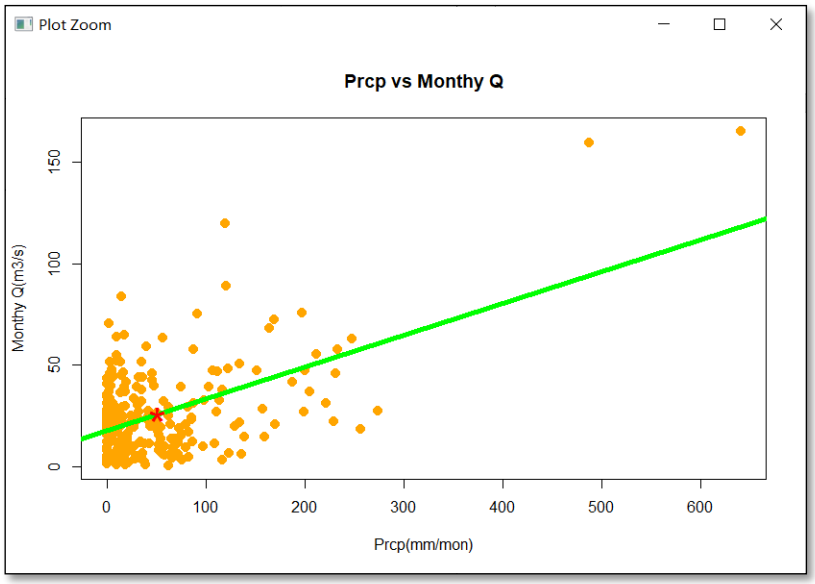


```
> summary(anova_one_way)
             Df Sum Sq Mean Sq F value Pr(>F)
prcp_diff     1    498   498.0   4.917 0.0273 *
Residuals   311  31497   101.3
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

7.3 探究逐月降雨总量对该流域的月平均流量的影响，二者是否存在线性相关关系？

因为流域的降雨量会对其流量产生影响，而逐日数据又不够直观，因此采用逐月的降雨总量和平均流量来进行分析，探究二者是否存在线性相关关系。

首先处理数据后得到逐月的降雨总量和平均流量数据，做出以月降雨量和月平均流量分别为 x，y 轴的散点图，再使用简单线性拟合，拟合结果如下所示，结果说明二者存在一定的线性相关关系。




```
> summary(fit4) #查看拟合结果，分析二者是否存在线性相关关系

Call:
lm(formula = data_liner$mean_q ~ data_liner$sum_prcp)

Residuals:
    Min      1Q  Median      3Q     Max
-39.600 -13.646  -1.723   9.606  83.076

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        17.97506    1.47948  12.150   <2e-16 ***
data_liner$sum_prcp  0.15596    0.01622   9.617   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.92 on 238 degrees of freedom
Multiple R-squared:  0.2798,    Adjusted R-squared:  0.2768
F-statistic: 92.48 on 1 and 238 DF,  p-value: < 2.2e-16
```