Question 1

1) Using the SATGPA data set in Stat2Data package. Test by using α= .01.

a. Create a new variable "SAT", which is the sum of MathSAT and VerbalSAT.

```
> install.packages("Stat2Data")
Installing package into '/Users/Yisha1/Library/R/3.5/library'
(as 'lib' is unspecified)
TRY URL'https://cran.rstudio.com/bin/macosx/el-capitan/contrib/3.5/Stat2Data_1.6.tgz'
Content type 'application/x-gzip' length 557245 bytes (544 KB)
==================================================
downloaded 544 KB



The downloaded binary packages are in
	/var/folders/10/15f3c4t54cvcrh62nnnwxcm00000gp/T//RtmpZ7pNS6/downloaded_pack
ages
> library(Stat2Data)
> data("SATGPA")
> SATGPA$SAT = SATGPA$MathSAT + SATGPA$VerbalSAT
> SATGPA$SAT
 [1] 1000 1200 1220 1270 1250 1130 1220 1190 1020 1200
[11] 1170 1260  650 1250 1080 1110 1260 1310 1380 1220
[21] 1230 1290 1250 1270
```

b. Create second new variable "SATLevel", and assign the value of "SATLevel" as 1 when SAT<=1100, 2 when 1100<SAT<=1200, 3 when 1200<SAT<=1300, and 4 when SAT>1300.

```
> SATGPA$SATLevel = numeric(nrow(SATGPA))
> SATGPA$SATLevel[SATGPA$SAT <= 1100] = 1
> SATGPA$SATLevel[1100 < SATGPA$SAT & SATGPA$SAT <= 1200] = 2
> SATGPA$SATLevel[1200 < SATGPA$SAT & SATGPA$SAT <= 1300] = 3
> SATGPA$SATLevel[SATGPA$SAT > 1300] = 4
> SATGPA$SATLevel
 [1] 1 2 3 3 3 2 3 2 1 2 2 3 1 3 1 2 3 4 4 3 3 3 3 3
```

c. Create third new variable "GPALevel" and assign the value of "GPALevel" as 1 when GPA<=2.8, 2 when 2.8<GPA<=3.3, 3 when 3.3<GPA<=3.5, and 4 when GPA>3.5

```
> SATGPA$GPALevel[SATGPA$GPA <= 2.8] = 1
> SATGPA$GPALevel[2.8 < SATGPA$GPA & SATGPA$GPA <= 3.3] = 2
> SATGPA$GPALevel[3.3 < SATGPA$GPA & SATGPA$GPA <= 3.5] = 3
> SATGPA$GPALevel[SATGPA$GPA > 3.5] = 4
> SATGPA$GPALevel
 [1] 2 2 2 2 4 1 1 2 1 2 2 2 2 4 2 2 2 2 1 4 1 2 3 4
```

d. Print out all the data in the ascending order of their GPALevel and the descending order of their SAT when GPALevel is the same.

```
> SATGPA2 = SATGPA[order(-SATGPA$GPALevel, SATGPA$SATLevel),]
> SATGPA2
   MathSAT VerbalSAT  GPA  SAT SATLevel GPALevel
5      620       630 3.61 1250        3        4
14     680       570 3.53 1250        3        4
20     670       550 3.53 1220        3        4
24     630       640 3.70 1270        3        4
23     600       650 3.50 1250        3        3
1      580       420 2.90 1000        1        2
13     350       300 3.13  650        1        2
15     550       530 3.10 1080        1        2
2      670       530 2.83 1200        2        2
8      690       500 3.00 1190        2        2
10     570       630 2.90 1200        2        2
11     620       550 3.00 1170        2        2
16     570       540 3.20 1110        2        2
3      680       540 2.90 1220        3        2
4      630       640 3.30 1270        3        2
12     690       570 3.25 1260        3        2
17     620       640 3.27 1260        3        2
22     590       700 3.30 1290        3        2
18     750       560 3.30 1310        4        2
9      520       500 2.77 1020        1        1
6      580       550 2.75 1130        2        1
7      620       600 2.75 1220        3        1
21     680       550 2.67 1230        3        1
19     700       680 2.60 1380        4        1
```

2) Use the Chi-Square test to conclude if the SATLevel and GPALevel are independent.

```
> t = table(SATGPA$SATLevel, SATGPA$GPALevel)
> chisq.test(t,correct = TRUE)

        Pearson's Chi-squared test

data:  t
X-squared = 7.4286, df = 9, p-value = 0.5926
```

Since p-value is 0.5926 > 0.01, fail to reject H0. The SATLevel and GPALevel are independent at 0.01 level.

3) Compute the mean and variance of “GPA” for each level of “GPALevel”, and

compute the correlation matrices for the four variables: MathSAT, VerbalSAT, GPA and SAT.

```
    > mean(SATGPA$GPA[SATGPA$GPALevel == 1])
    [1] 2.708
    > var(SATGPA$GPA[SATGPA$GPALevel == 1])
    [1] 0.00512
    > mean(SATGPA$GPA[SATGPA$GPALevel == 2])
    [1] 3.098571
    > var(SATGPA$GPA[SATGPA$GPALevel == 2])
    [1] 0.0303978
    > mean(SATGPA$GPA[SATGPA$GPALevel == 3])
    [1] 3.5
    > var(SATGPA$GPA[SATGPA$GPALevel == 3])
    [1] NA
    > mean(SATGPA$GPA[SATGPA$GPALevel == 4])
    [1] 3.5925
    > var(SATGPA$GPA[SATGPA$GPALevel == 4])
    [1] 0.006558333
    > cor(SATGPA[,1:4])
                 MathSAT VerbalSAT       GPA       SAT
    MathSAT    1.0000000 0.5103260 0.0543507 0.8584501
    VerbalSAT  0.5103260 1.0000000 0.2444543 0.8791712
    GPA        0.0543507 0.2444543 1.0000000 0.1759089
    SAT        0.8584501 0.8791712 0.1759089 1.0000000
```

4) Do the data provide sufficient evidence to indicate that the mean of MathSAT is significantly greater than the mean of VerbalSAT

```
> diff = SATGPA$MathSAT -  SATGPA$VerbalSAT
> shapiro.test(diff)

	Shapiro-Wilk normality test

data:  diff
W = 0.95673, p-value = 0.3763

> t.test(diff, alternative = "greater")

	One Sample t-test

data:  diff
t = 3.2059, df = 23, p-value = 0.001961
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 25.0156      Inf
```

```
sample estimates:
mean of x
    53.75
```

Since

Since p-value for shapiro test is 0.3763 > 0.01, data is normal. We can use t test. P-value for t test is 0.00191< 0.01. We can reject null. Therefore, mean of MathSAT is significantly greater than the mean of VerbalSAT at 0.01 level.

5) Test if the proportion of MathSAT less than VerbalSAT is 0.5.

```
> x = sum(SATGPA$MathSAT > SATGPA$VerbalSAT)
> n = nrow(SATGPA)
> x
[1] 17
> n
[1] 24
> prop.test(x, n, 0.5)

    1-sample proportions test with continuity
    correction

data:  x out of n, null probability 0.5
X-squared = 3.375, df = 1, p-value = 0.06619
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.4875243 0.8656176
sample estimates:
        p
0.7083333

> binom.test(x, n, 0.5)

    Exact binomial test

data:  x and n
number of successes = 17, number of trials = 24,
p-value = 0.06391
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.4890522 0.8738479
sample estimates:
probability of success
             0.7083333
```

We can us prop test or binom test for both cases. Both p-values are larger than 0.01, we can’t reject H0 that the proportion of MathSAT greater than VerbalSAT is 0.5. Therefore, he proportion of MathSAT greater than VerbalSAT is 0.5 at 0.01 level.

Question 2

Analyze and interpret the effect of explanatory variables on the milk intake (dl.milk) in the kfm data set (ISwR) using a multiple regression model. Test by using α= .05.

1) Run regression for dl.milk on all other variables. Do you find any significance that milk intake can be explained by other variables?

```
> library(ISwR)
> data(kfm)
> l = lm(dl.milk ~ ., data = kfm)
> summary(l)

Call:
lm(formula = dl.milk ~ ., data = kfm)

Residuals:
     Min       1Q   Median       3Q      Max
-1.89286 -0.87720  0.06426  0.73663  2.28685

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -11.652909   4.357128  -2.674 0.010542 *
no           -0.005522   0.005286  -1.045 0.302010
sexgirl      -0.488757   0.312518  -1.564 0.125164
weight        1.311822   0.324088   4.048 0.000212 ***
ml.suppl     -0.002432   0.001254  -1.939 0.059077 .
mat.weight    0.002453   0.023956   0.102 0.918925
mat.height    0.076445   0.030401   2.515 0.015739 *
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.074 on 43 degrees of freedom
Multiple R-squared:  0.5571,   Adjusted R-squared:  0.4953
F-statistic: 9.015 on 6 and 43 DF, p-value: 2.189e-06
```

Since p-value is 2.189e-06 < 0.05, the significant variables which can explain milk intake are weight and mat height at 0.05 level.

2) Find regression models in which fewer explanation variables should be used. i.e., select a subset of variables so that a better fit can be achieved.

```
> l2 = lm(dl.milk ~ weight + mat.height, data = kfm)
> summary(l2)

Call:
lm(formula = dl.milk ~ weight + mat.height, data = kfm)

Residuals:
     Min       1Q   Median       3Q      Max
-2.19598 -0.82149  0.01822  0.75582  2.83375

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -11.92014    4.07325  -2.926  0.00527 **
weight        1.42862    0.31338   4.559 3.67e-05 ***
mat.height    0.07063    0.02636   2.680  0.01013 *
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.109 on 47 degrees of freedom
Multiple R-squared:  0.4835, Adjusted R-squared:  0.4615
F-statistic:    22 on 2 and 47 DF,  p-value: 1.811e-07
```

Both variables are still significant at 0.05 level and the model has a R-squared which is slighter lower than full model, since it has much smaller size. It should be fitted better.