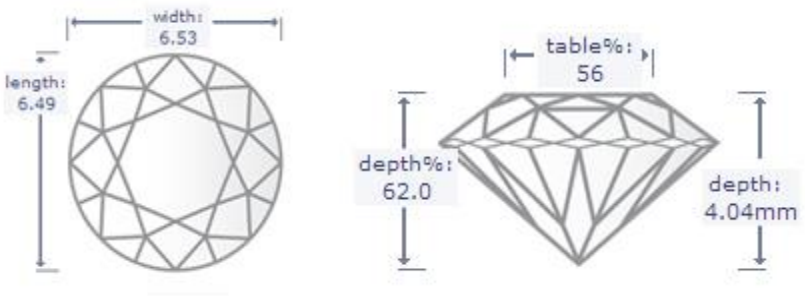
Name:

Exam IOU35A Univariate Data and Modelling
Prof. A. Carbonez
Tuesday, August 16 2016, 9.00
Data: diamonds2.txt
This data frame contains measurements on 5000 diamonds.

Variable name	Description
Price	Price diamond, US $
Carat	Weight of a diamond
Cut_new	Quality (of the cut) of the diamond (from less good to very good) : (Fair, Good, Very Good, Premium, Ideal)
Color	Colour of the diamond (from less good to very good): (J, ….,D)
Clarity	Purity of the diamond. (from less good to excellent purity): (I1, SI1, SI2, VS1, VS2, VVS1, VVS2, IF)
x	Length of the diamond, in mm (length)
y	Width of the diamond, in mm (width)
z	Depth of the diamond, in mm (depth)
depth	Total depth percentage
table	Width of the top of diamond relative to the max. width of the diamond
heavy	0: diamond is not so heavy 1: diamond is heavy



Name:

1. **diamonds2**

Here we are only interested in diamonds with quality (cut_new) *Fair, Very good or Ideal*.

For these diamonds, we want to check whether there exists a significant difference in average price between diamonds of different quality (**use *cut_new***).

Be complete and give your answer in a structured way! Formulate clearly the hypothesis you make (use symbols !). **Use here significance level 0.05**. Interpret your result in terms of the concrete question.
If there are significant differences in average price between diamonds of different quality, give more information about which quality of diamonds give significant different average prices than other quality of diamonds. Formulate your result.

Name:

2. diamonds2 (for a description of the variables, take a look at the first page)
Here we only use the diamonds with the best clarity (clarity = 'IF')

Use a linear regression model for the price of diamonds where we use as explanatory variables the length of the diamond (x), whether a diamond is heavy or not (heavy) and the interaction term.

In the output below: clar_P= Price, clar_H: heavy, clar_x: x,

Use significance level 0.10 to answer following questions.

```
Call:
lm(formula = clar_P ~ clar_H + clar_x + I(clar_x * clar_H))

Residuals:
    Min      1Q  Median      3Q     Max
-981.04 -196.73  -40.86  153.65 1166.51

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -5838.86     421.69 -13.846   <2e-16 ***
clar_H             1958.73   12544.08   0.156    0.876
clar_x             1613.30      78.04  20.672   <2e-16 ***
I(clar_x * clar_H)  -385.68    2077.60  -0.186    0.853
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 404.1 on 89 degrees of freedom
Multiple R-squared:  0.8324,    Adjusted R-squared:  0.8267
F-statistic: 147.3 on 3 and 89 DF,  p-value: < 2.2e-16
```

a. Specify the population regression model (starting model) for the price of a diamond. (short answer)

Name:

b. Is the effect of x on the average price of a diamond the same for diamonds which are heavy and for diamonds which are not so heavy? Give H0 and H1 to test this and interpret your result.

c. Consider now the linear regression model for the price of diamond when we have as explanatory variables the length (clar_x) and whether diamond is heavy or not (clar_h).

```
Call:
lm(formula = clar_P ~ clar_H + clar_x)

Residuals:
   Min     1Q Median     3Q    Max
-981.5 -196.5  -46.2  153.6 1166.3

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -5835.93     419.13 -13.924   <2e-16 ***
clar_H       -369.67     191.79  -1.927   0.0571 .
clar_x       1612.75      77.57  20.792   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 401.9 on 90 degrees of freedom
Multiple R-squared:  0.8323,    Adjusted R-squared:  0.8286
F-statistic: 223.3 on 2 and 90 DF,  p-value: < 2.2e-16
```

Answer following questions.
Has the heaviness of a diamond an effect on the price of the diamond, taking into account the length of the diamond?
Specify H0 and H1 you need to answer this question. Give the obtained p-value and formulate your conclusion in terms of this concrete question.

Name:

d. **Retake the output given in question c. Is there a different regression model (for the price of a diamond in term of the length of the diamond) for diamonds which are heavy and for diamonds which are not heavy?**

✓ **In case your answer is NO. Give here the estimated regression model for diamonds in terms of length of the diamond.**

✓ **In case your answer is YES:**

(i) Estimated regression model for diamonds in terms of length for heavy diamonds:

(ii) Estimated regression model for diamonds in terms of length for non-heavy diamonds:

Name:

3. Diamonds2 (for a description of the variables, take a look at the first page)

Part 3.1.
Check whether the average width of diamonds (y) is significant larger than the average length of diamonds (x). Check this separately for diamonds with the best color (D) and for diamonds with the worst color (I). Use each time significance level 0.05.

Be complete in your answer. Formulate H0 and H1 whenever necessary. Visualize the obtained p-values and interpret the obtained result in terms of the formulated question.

Name:

Part 3.2.We only consider here diamonds with the worst color (I)

a. **What is the power of the previous test when the true difference between width (y) and length (x) is 0.005. You can assume a standard deviation of 0.05.**

b. **What is the sample size needed to obtain a power of at least 0.80 in the situation of a?**

Name:

4.**We now consider only two groups of diamonds. The diamonds with color D and the diamonds with color E.** Check whether the proportion of heavy diamonds (heavy=1) is the same in both groups.

Name:

5. Short questions. **Be careful, these are multiple choice questions with GIS correction. If your answer is correct you'll get 0.5. If your answer is incorrect, you lose 0.25.**

5.1. For one dataset, we fit 3 different logistic models. Some results are given in het table below. Can you say which model will have the highest proportion of deviance explained by the model?

	model	*Residual deviance*
Model 1	$\hat{p}' = \beta_0 + \beta_4 X_4$	*102*
Model 2	$\hat{p}' = \beta_0 + \beta_1 X_1 + \beta_2 X_2$	
Model 3	$\hat{p}' = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$	*100*

a. Model 1
b. Model 2
c. Model 3
d. we do not have enough information to decide this

Name:

5.2. We consider a logistic regression model for being a heavy diamond or not (we use the variable heavy) in function of the price of the diamond. We give below the results of a simple logistic regression model.

```
> glm.log1 <- glm(heavy ~ price, family=binomial(link=logit),
data=diamonds_2)
> summary(glm.log1)

Call:
glm(formula = heavy ~ price, family = binomial(link = logit),
    data = diamonds_2)
```

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.6385206  0.3688922  -23.42   <2e-16 ***
price        0.0026670  0.0001142   23.35   <2e-16 ***
---
```

Consider now two diamonds. Diamond 1 has a price Z. Diamond 2 has a price 100+Z. What can you say about the odds for being a heavy diamond?

(i) Odds (Diamond2) = 0.76 * Odds(Diamond1)
(ii) Odds(Diamond2) = -8.33 * Odds(Diamond1)
(iii) Odds(Diamond2) = 1.3 * Odds(Diamond1)
(iv) No one of the above