Problem 1 (6 credits)

0
1
2

a)

Question N NY1

The olive ckage contains the % of 8 fatty acids found in Italian olive oils.

```
library(da
data(oliv
head(oliv
```

```
##                                palmitic palmitoleic stearic oleic linoleic
## 1 Southern Italy North-Apulia    10.75        0.75    2.26 78.23     6.72
## 2 Southern Italy North-Apulia    10.88        0.73    2.24 77.09     7.81
## 3 Southern Italy North-Apulia     9.11        0.54    2.46 81.13     5.49
## 4 Southern Italy North-Apulia     9.66        0.57    2.40 79.52     6.19
## 5 Southern Italy North-Apulia    10.51        0.67    2.59 77.71     6.72
## 6 Southern Italy North-Apulia     9.11        0.49    2.68 79.24     6.78
##   linolenic arachidic eicosenoic
## 1      0.36      0.60       0.29
## 2      0.31      0.61       0.29
## 3      0.31      0.63       0.29
## 4      0.50      0.78       0.35
## 5      0.50      0.80       0.46
## 6      0.51      0.70       0.44
```

Write R code using ggplot2 to plot the distribution of the % of the fatty acids except oleic using density plots with different line colors to distinguish each fatty acid. Give meaningful labels to both axes.

```
dt <- as.data.table(olive)
dt <- dt[, oleic := NULL]
oliver_dt <- melt(dt, id.vars= c("region", "area") variable.name = "fatty_acid",
value.name = "value")
oliver_dt
ggplot(olive_melted, aes(x=variable, color = fatty_acid)) +
 geom_density()
```

0
1
2

b)

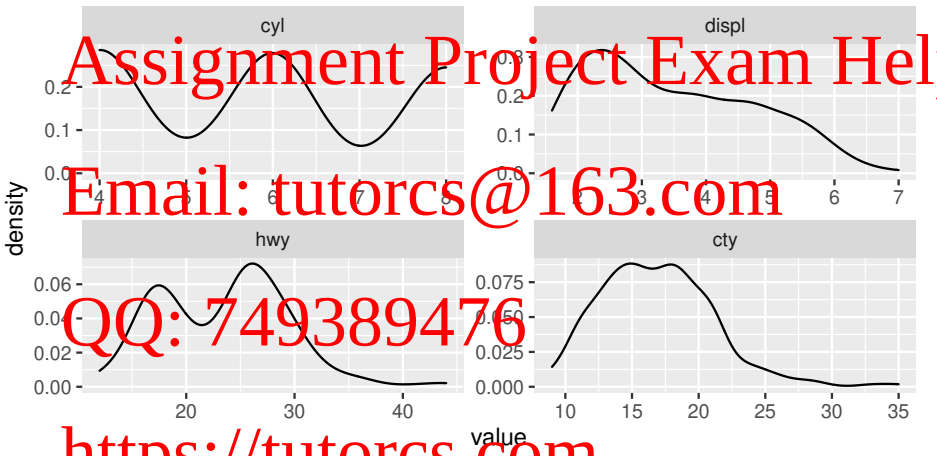
Question Nr. 3QM41LNA09GZ73DB55AQ5

The mpg da kage contains different cars features, mostly involving fuel.

```
library(g
data(mpg)
head(mpg)
```

```
## # A til
##   manu        ar   cyl trans      drv     cty   hwy fl    class
##   <chr>       <int> <int> <chr>      <chr> <int> <int> <chr> <chr>
## 1 audi        1999     4 auto(15)   f        18    29 p     compa~
## 2 audi   a4  1.8  1999     4 manual(m5) f        21    29 p     compa~
## 3 audi   a4  2    2008     4 manual(m6) f        20    31 p     compa~
## 4 audi   a4  2    2008     4 auto(av)   f        21    30 p     compa~
## 5 audi   a4  2.8  1999     6 auto(15)   f        16    26 p     compa~
## 6 audi   a4  2.8  1999     6 manual(m5) f        18    26 p     compa~
```

Write R code that produces the following plot.




```
data <- as.data.table(mpg)
data <- melt(data, id.vas = c("manufacturer", "model", "year", "trans", "fl", "class"))
```

0
1
2

c)

Question Nr. 0GW52IK81LNA06UY23CW8

The admiss s package provides the number of applicants and admitted students to 6 different Write R code using `ggplot2` to plot the difference between applicants and admitt using bars and stratified by gender using facets. Give meaningful labels to bo obtain negative values.

```
library(da
data(admi
admission
```

```
##    maj      cants
## 1      A    825
## 2      B    560
## 3      C    325
## 4      D    417
## 5      E    191
## 6      F    373
## 7      A    108
## 8      B     25
## 9      C    593
## 10     D    375
## 11     E    393
## 12     F    341
```

Problem 2 (2 credits)

0 1 2

Which operation has been applied to table A and table B to return the result table? Justify your answer. Write one line of R code that would produce the result table assuming a data table A and a data table B in the working environment.

Table A:

	ard	CCV	type
	746651971	582	l
	911423246	221	i
	899774435	142	r
	946592947	479	l
	521545978	881	l
	281216234	698	o
8	1764682661721638	566	o
24	2622321425978251	528	o
19	7271112241595296	393	o
18	4225693846619738	421	r

Table B:

firstName	lastName	customer_id
Aamina	el-Sinai	16
Marcus	Hendrix	13
Derek	Martinez	8
Muntasir	al-Sharifi	24
Alexis	Smith	19
Alexis	Arreola	18
Khanea	Forrest	1
Julia	Deronde	6
Keith	Hart	25
Tiana	Ramirez	9
Adam	Highman	2

Result table:

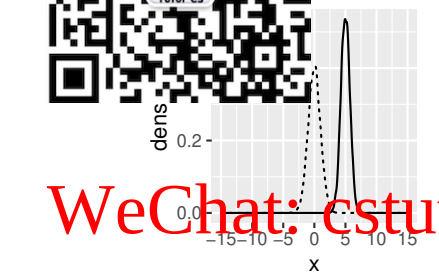
id	CreditCard	CCV	type	firstName	lastName
1	NA	NA	NA	Khanea	Forrest
2	NA	NA	NA	Adam	Highman
6	NA	NA	NA	Julia	Deronde
8	1764682661721638	566	o	Derek	Martinez
9	NA	NA	NA	Tiana	Ramirez
13	3923818281216234	698	o	Marcus	Hendrix
14	7393954899774435	142	r	NA	NA
15	1837655746651971	582	l	NA	NA
16	7364376521545978	881	l	Aamina	el-Sinai
18	4225693846619738	421	r	Alexis	Arreola
19	7271112241595296	393	o	Alexis	Smith
21	5927428911423246	221	i	NA	NA
23	7844437946592947	479	l	NA	NA
24	2622321425978251	528	l	Muntasir	al-Sharifi
25	NA	NA	NA	Keith	Hart

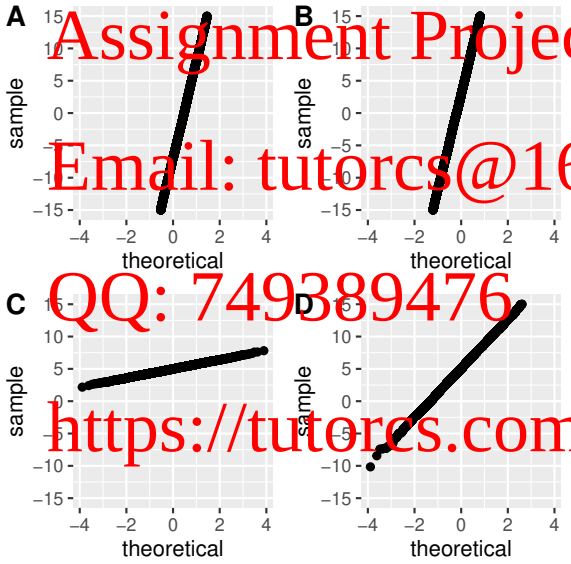
Problem 3 (2 credits)

a) 0 1

Question Nr. 6CK12JZ6C:

Which normal Q-Q plot (i rd Normal distribution) A, B, C, or D corresponds to the distribution depicte he standard Normal distribution, i.e. the Gaussian distribution with mean 0 a plot below with a dashed line.



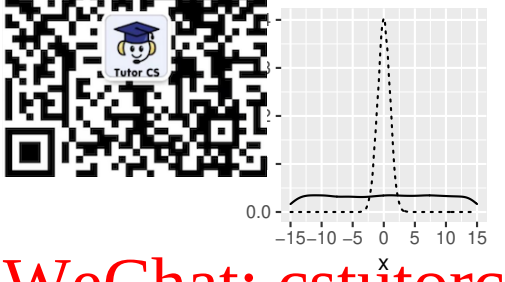


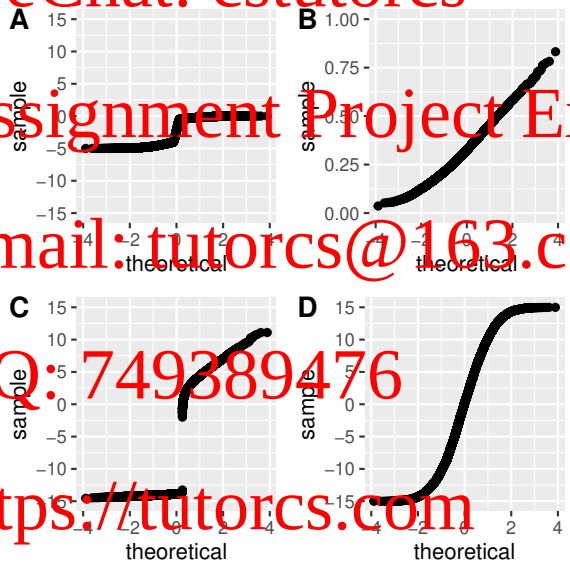
0
1

b)

Question Nr. 3JN34PA3DD3YU35RP0

Which norma st the standard Normal distribution) A, B, C, or D corresponds to the distrib lot below? The standard Normal distribution, i.e. the Gaussian distribution v shown in the plot below with a dashed line.





Problem 7 (2 credits)

a) 0 1

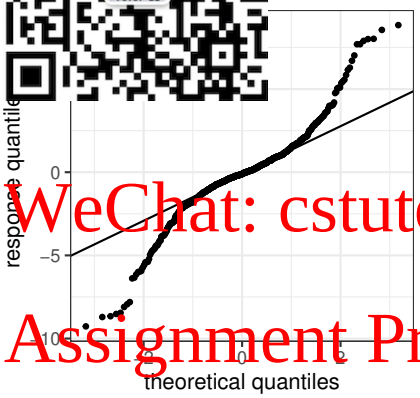
Question Nr. 8ZF48OQ3WP45KH55GQ0

We consider a linear regression model parameterized as

ϵ_i

where $i = 1...N$ denotes the data point i ble, α and β the coefficients, x_i the explanatory variable and ϵ_i the error term. Let $\hat{y}_i$ be the fitt

Does the following plot provide evidenc e linear regression? Justify.



0
1

b)

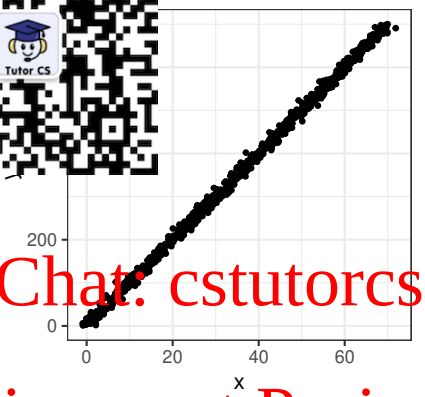
Question Nr. 0JU8SP106HN41KM58HL9

We consider a linear regression model parameterized as

$$y_i = \alpha + \beta \cdot x_i + \epsilon_i$$

where $i = 1...N$ denotes the data point indices, y_i is the response variable, α and β the coefficients, x_i the explanatory variable and ϵ_i the error term. Let

Does the following plot pr mptions of the linear regression? Justify.



Problem 8 (2 credits)

0
1
2

Question Nr. 4LY90YX24AQ71F66EX2

```
library(dslabs)
```

Consider the "brca" dataset from [illegible]stic regression model which predicts the response variable brca$y given the feature perimeter_se. As[illegible] of the logistic regression model are met. Starting from an original probability of 10 % of malignant ([illegible]e probability of developping a malignant (cancer) increase, when the feature perimeter_se increases

Problem 9 (2 credits)

0
1
2

QuestionId: 4YF2RH7TD00JE84ZN0

Consider the features smoothness_mean, radius_mean from the brca dataset. Provide R code that plots a ROC curve of both features as predictors of malignancy (variable bres$y == "M"), and indicate the feature that has the highest true positive rate when the false positive rat

Problem 10 (2 credits)

0
1
2

```
library(dslabs)
library(data.table)
```

Question Nr. 4IL88XL37CT72GN

Consider the variable 'concavity_w A researcher states that no other variable from the matrix 'brca$x' associates with the variable 'conc Spearman's correlation. Do you reject this hypothesis using the significance level of 1%? Provide e er. Do not mind warnings, if any, about exact p-values with ties.

Problem 11 (1 credit)

0
1

Quest[illegible]V8LMHZ4060

Assume 5[illegible]irst clustering gave the clusters {v} and {w,x,y,z}. One then run [illegible]ers: {v,w,x} and {y,z}. Applying hierarchical clustering yields als[illegible]z}. Which of the k-mean clustering and the hierarchical clustering[illegible]nilar to the first clustering? Support your answer with a metric lea[illegible]