Example exam

Welcome to the Practice Exam. The exam consists of two parts: a **theoretical part** and a **practical part**.

You have XX hours in total. You have to upload your answers on blackboard before the deadline (XX:XX). If you have a valid reason for an extension, you are allowed to submit until 30 minutes after the deadline.

Making the exam

- Open the project file `Practice.Rproj` in RStudio before you start.
- Your answers to both the theoretical and practical parts will be in R markdown format. Use the prepared answer file `Practice_0000000.Rmd`, replace `0000000` with your student number and also put your student number in the `Author` field at the top of the file. See the next section for hand-in details.
- The exam is "open book": you can use the internet, you can use your notes, you can use the reading materials, you can use the lecture slides, etc.
- You are **not allowed to communicate with others in any way**. We trust that you abide by this rule. We will use various methods to check for fraudulent entries.

Handing in the exam

You will hand in a zipped folder with the following files on blackboard:

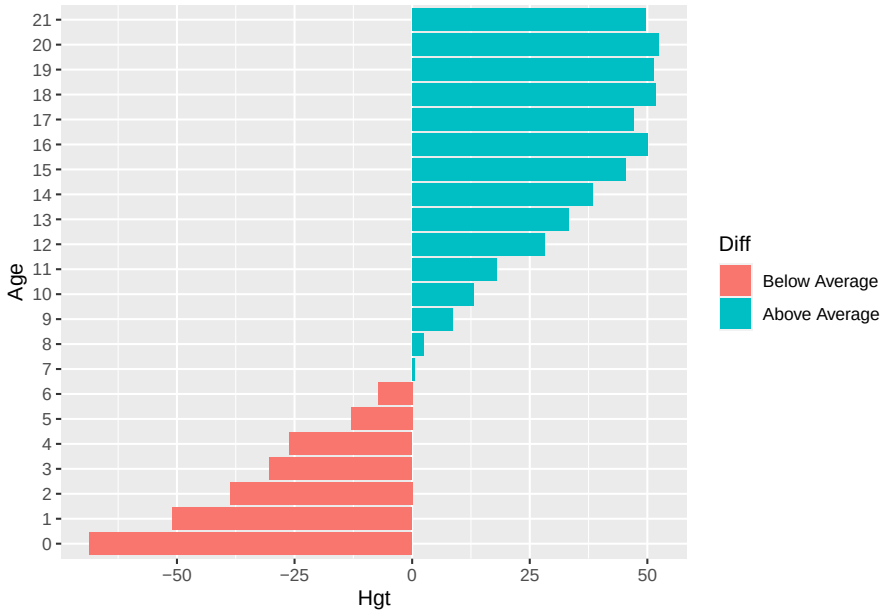
- `Practice_0000000.Rmd` (replace 0000000 with your student number)
- `/data` (folder with data to which the .Rmd file refers)
- `Practice_0000000.html` (Compiled version of your answer R markdown. Please try to submit this but if compiling does not work don't worry about it!)

Make sure to give yourself enough time at the end for zipping and uploading.

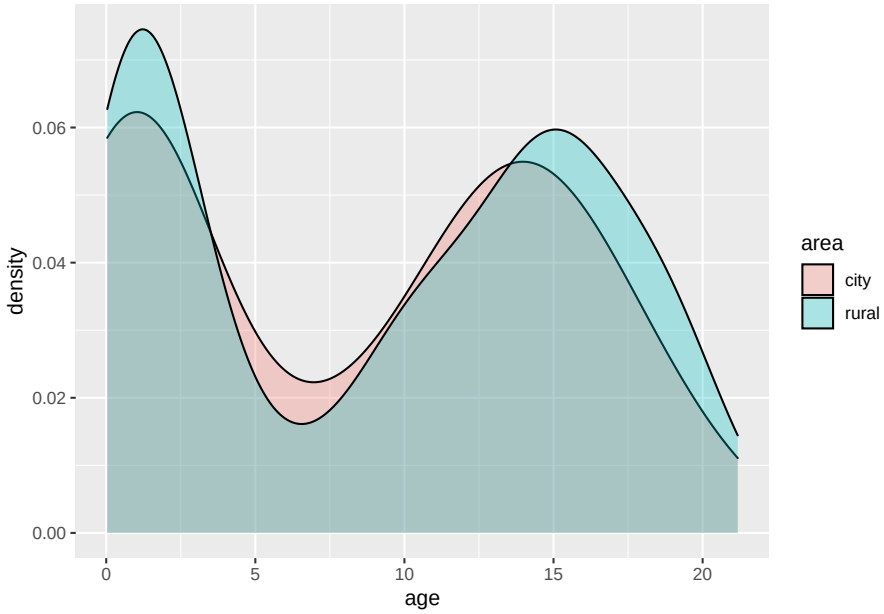
Theoretical part [10 points]

Q1 Data Visualization

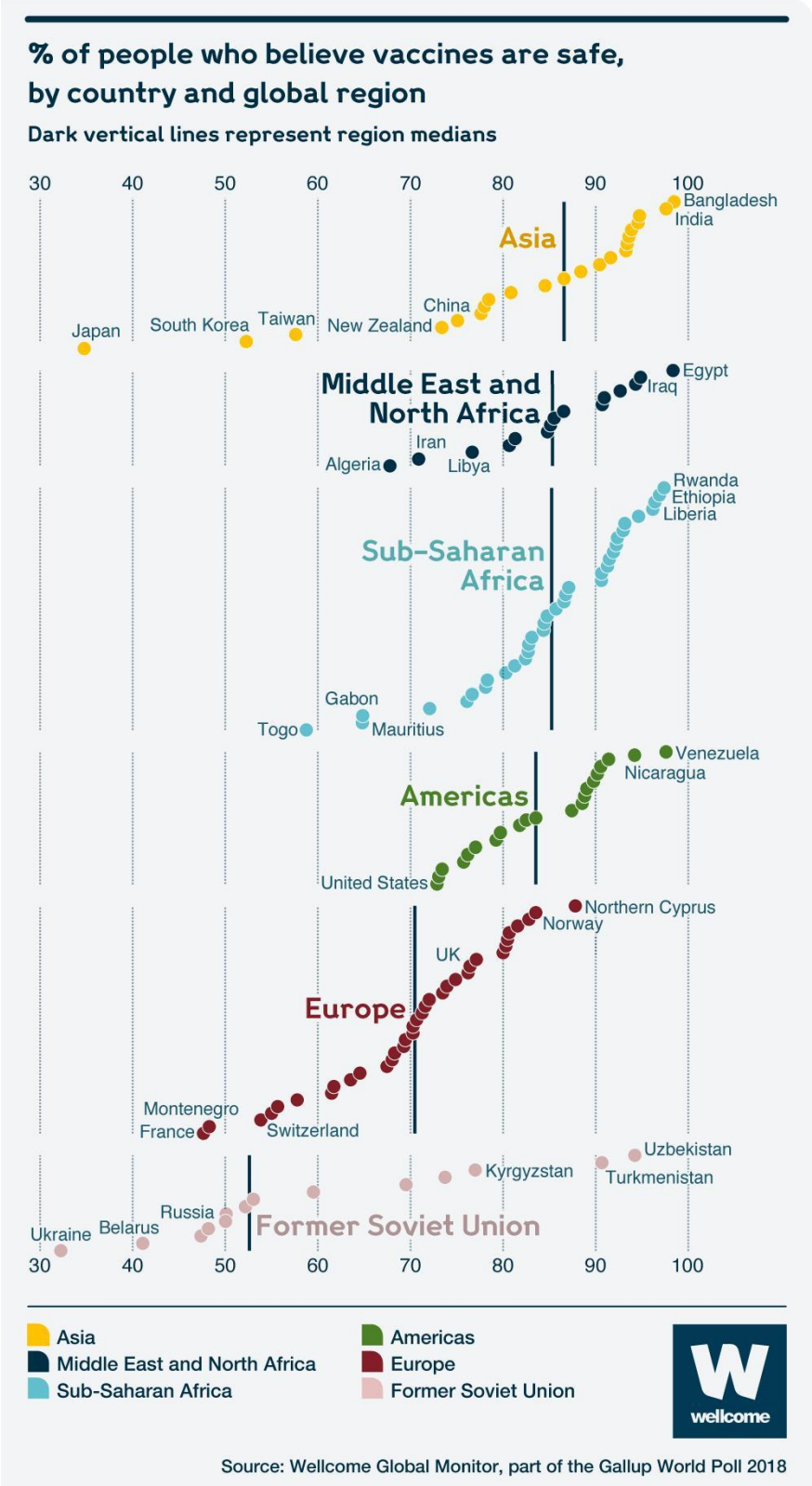
a) For the following plot, name the aesthetics (that is, name the mapping of variables to aesthetics), geoms, and scales. If applicable, name any facets, transformations, or special coordinate systems.



b) For the following plot, name the aesthetics (that is, name the mapping of variables to aesthetics), geoms, and scales. If applicable, name any facets, transformations, or special coordinate systems.



Q2 Give 3 suggestions for improvement of the following plot, and explain your rationale based on visualization principles. If it's too small, you can zoom in on to the pdf.



Q3 Energy prediction You work for an energy company. From the cool smart energy meters in every customer's home you can collect features, measured on 15 minute intervals. Your goal is to predict whether the energy usage is over their prepayment or under (i.e. too much used or too little used. You have a hand-coded label for each row in the data. Your dataset has 5000 columns and 2000 rows.

a) You want to perform logistic regression but it does not work when you have more columns than rows. What would your strategy be? Be specific about the steps you would take!

b) From your logistic regression model, you obtain the confusion matrix below. What is the accuracy? Is this high compared to the baseline accuracy?

```
##           true
## predicted over under
##     over  1752    26
##     under  148    74
```

Practical part [11 points]

Load the following packages:

```
library(DAAG)
library(glmnet)
```

Q5 Decision rule

a) In the data set `head.injury` (from package `DAAG`), obtain a logistic regression model relating `clinically.important.brain.injury` to all the other variables.

b) Patients whose risk is sufficiently high will be sent for CT (computed tomography). Using a risk threshold of 0.025 (2.5%), turn the result into a decision rule for use of CT and indicate three different scenarios that would satisfy the threshold.

Q6 LASSO

In this question, the goal is to predict `y` from `x`.

a) Load the workspace `data.Rdata` and show an informative plot of the `y` and `x` space.

b) Run a 10-fold cross-validated LASSO logistic (`family = "binomial"`) regression using the misclassification error as the criterion.

c) Make a prediction of the class labels at $\lambda = 0.05, 0.01$

d) create a plot that shows the values of λ. What is the optimal value and why?