Data Science Concepts – Final Exam

Make-Up questions for the final exam

First Name, Last Name: ____________________________________

Matric. No.: ____

Signature: ____

Instructions

- Whenever you see lines "_________" or checkboxes □ please fill in the blanks or check the boxes within the text.
- Whenever you see **(ANSWER ON BACKSIDE/EXTRA-SHEET)** please write your answer on the backside of the sheet or on an extra sheet starting with the question number and subnumbers.

THIS IS A MOCK EXAM! In the real exam

- there will also be questions on different topics
- you will know the total points and the points per question before
- you will be graded by the percentage of points achieved and there will be 100 points in total
- the points you achieved will be scaled up by a factor of 1.2
  - That means for example: 50 points count as 60, 80 points count as 96. So, you do not need to solve all questions to achieve the best grade.
  - The factor 1.2 may be adjusted to account for the difficulty of the exam.

Questions

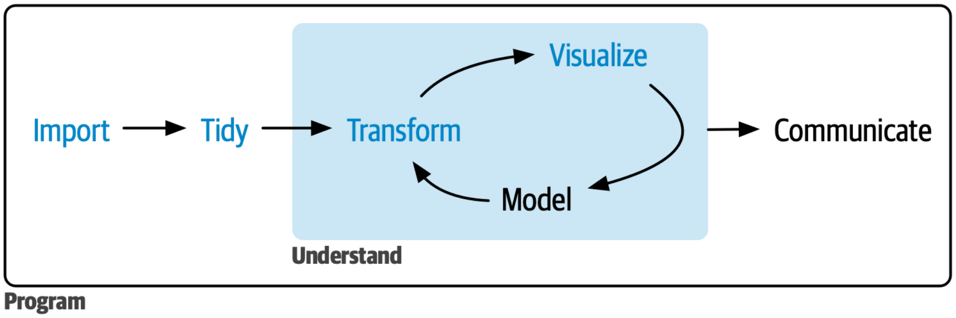
1. We define the tools of Data Science, Machine Learning, and Artificial Intelligence by the main outcomes they produce: *Data Science* (DS) produces insights, *Machine Learning* (ML) produces predictions, and *Artificial Intelligence* (AI) produces decisions.

   a. Based on this definition, how would you classify the following practices as *DS*, *ML*, or *AI*?

      i. A movie and TV show streaming company analyzes the customer satisfaction with their recommendation algorithm. _____
      ii. A mobile phone company has a customer churn model which predicts the likelihood that a customer will cancel the contract in the next months. _____
      iii. A fingerprint recognition system on a smartphone decides if a certain person can log in to use the smartphone. _____
      iv. An insurance company uses a model and geo-referenced geological and meteorological data to quantify the risk of damage from natural disasters for every building insured. _____
      v. Political analysts try to understand the main causes for the results of a referendum being party affiliation, gender, age, or the place of residence of voters. _____
      vi. A chatbot answers questions which humans write in a chatbox on a website. _____

   b. **(ANSWER ON BACKSIDE/EXTRA-SHEET)** Think of a company running an online dating application where users can search for, contact, and chat with potential dating partners. Give a brief example of potential applications of Data Science, Machine Learning, and Artificial Intelligence of that company.

2. This is a visualization of the data science process.



   Assing the terms *Import*, *Tidy*, *Transform*, *Visualize*, *Model*, *Communicate* to the following activities.

   a. Creating a variable `year` and `gdp` in a data frame from the columns `year_1950`, `year_1951`, …, `year_2022` which have values for GDP.

____

b. Creating a new variable `log_gdp` by taking the logarithm of the variable ‘gdp’.

____

c. Loading a data frame from a csv file.

____

d. Rendering a report from a markdown file showing the results of a growth analysis.

____

e. Creating a scatter plot of the variables `log_gdp` and `year`.

____

f. Fitting a line with `log_gdp` as dependent and `year` as independent variable.

____

3. *Date frames.* Which of the following is correct?

   □ A data frame can store data vectors of different types as variables.

   □ A data frame can store data vectors of different lengths as variables.

4. When you create a vector with entries of different data types, types will be coerced. What will be the output of the vector created with `c(3, TRUE, 4, FALSE)`?

   □ `[1] 3 0 4 1`

   □ `[1] 3 1 4 0`

   □ `[1] 3 4`

   □ `[1] TRUE TRUE TRUE FALSE`

5. The vector `c(160,200,240,250,250,400,600)` is the vector of 7 incomes of a group of people. Which of the following statements is correct (there maybe multiple or no correct answer!).

   □ The median income is 250

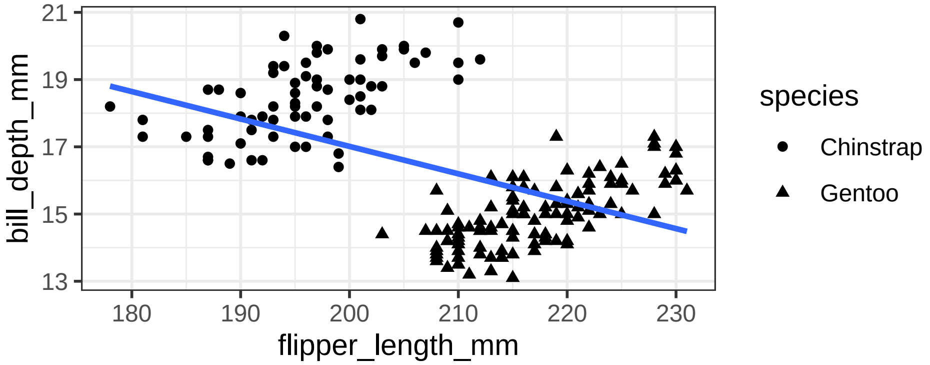
   □ The mean income is 600

   □ By adding an 8th value of your choice you can create a vector which median is 400

   □ By adding an 8th value of your choice you can create a vector which mean is 400

6. **(ANSWER ON BACKSIDE/EXTRA-SHEET)** Is it useful to split the data into training and testing data for (a) a predictive question, or (b) an explanatory inferential questions? Describe for both either for what purpose it is useful or why it is not necessary.

7. **(ANSWER ON BACKSIDE/EXTRA-SHEET)**
This scatter plot is made using the `penguins` data frame filtered for the Chinstrap and Gentoo penguins only.



Model 1: Below is the output of the corresponding linear model for `bill_depth_mm ~ flipper_length_mm`.

```
# A tibble: 2 x 5
  term              estimate std.error statistic  p.value
  <chr>                <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)        33.3      2.09         16.0  1.28e-36
2 flipper_length_mm  -0.0816   0.00995     -8.19 4.06e-14
```

The `r-squared` value of *Model 1* is 0.266306.

Model 2: Below is the output of the model with interaction effects of species `bill_depth_mm ~ flipper_length_mm * species`.

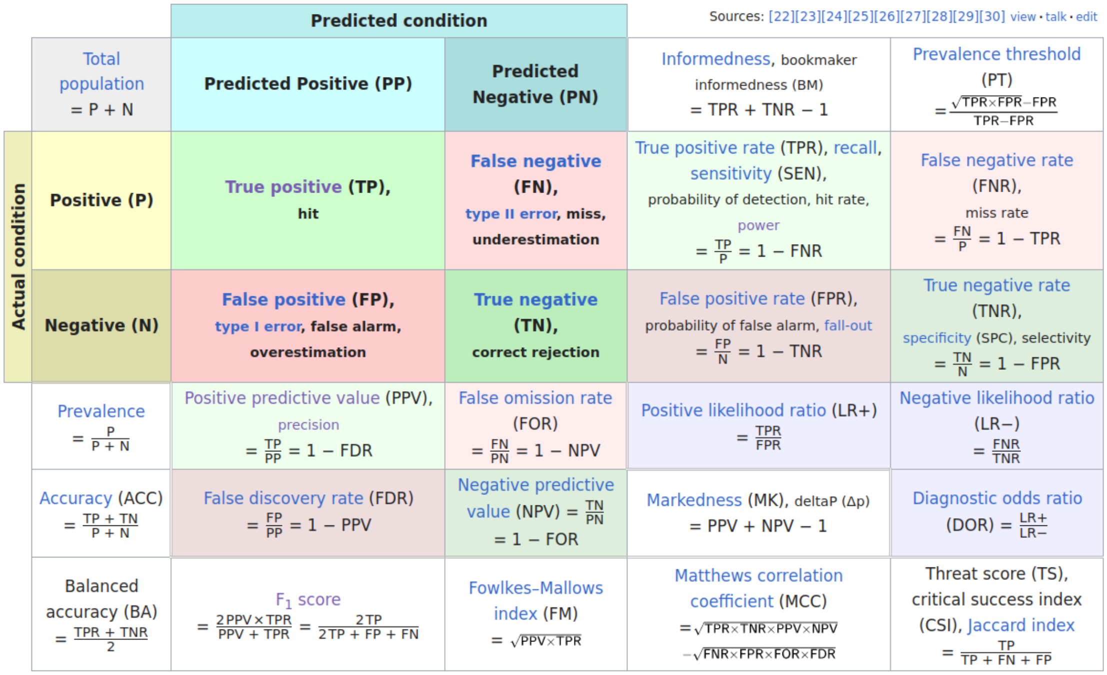
```
# A tibble: 4 x 5
  term                             estimate std.error statistic  p.value
  <chr>                               <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)                        0.335     2.65        0.126 9.00e- 1
2 flipper_length_mm                  0.0924    0.0135      6.83  1.22e-10
3 speciesGentoo                     -8.45      3.57       -2.36  1.91e- 2
4 flipper_length_mm:speciesGentoo    0.0140    0.0175      0.804 4.22e- 1
```

The `r-squared` value of *Model 2* is 0.8389515.

a. For *Model 1*: Interpret intercept, coefficient and R-squared! What does the intercept mean quantitatively? What does the coefficient tell us quantitatively about the relation between flipper length and bill depth? What percentage of the variance in bill depth is explained by flipper length? Is it a good model fit?

b. For *Model 2*: Which species has been automatically chosen as the reference category? The model describes two lines, one for each species, each with an own intercept and slope. Explain how we can infer intercept and slope for each species? Explain for each species how a mm in flipper length changes bill depth on average! Sketch the two lines in the plot above.
c. Compare the variance explained of the new model with the model without species. Which model is better?
d. In comparison of coefficients of *Model 1* and *Model 2*, does something appear paradoxical? If yes, explain? How is this "paradox" called?

8. Using the *confusion matrix*



Sources: [22][23][24][25][26][27][28][29][30] view • talk • edit

		Predicted condition			
	Total population $= P + N$	Predicted Positive (PP)	Predicted Negative (PN)	Informedness, bookmaker informedness (BM) $= TPR + TNR - 1$	Prevalence threshold (PT) $= \frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$
Actual condition	Positive (P)	**True positive (TP),** hit	**False negative (FN),** type II error, miss, underestimation	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{TP}{P} = 1 - FNR$	False negative rate (FNR), miss rate $= \frac{FN}{P} = 1 - TPR$
	Negative (N)	**False positive (FP),** type I error, false alarm, overestimation	**True negative (TN),** correct rejection	False positive rate (FPR), probability of false alarm, fall-out $= \frac{FP}{N} = 1 - TNR$	True negative rate (TNR), specificity (SPC), selectivity $= \frac{TN}{N} = 1 - FPR$
	Prevalence $= \frac{P}{P + N}$	Positive predictive value (PPV), precision $= \frac{TP}{PP} = 1 - FDR$	False omission rate (FOR) $= \frac{FN}{PN} = 1 - NPV$	Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$	Negative likelihood ratio (LR−) $= \frac{FNR}{TNR}$
	Accuracy (ACC) $= \frac{TP + TN}{P + N}$	False discovery rate (FDR) $= \frac{FP}{PP} = 1 - PPV$	Negative predictive value (NPV) $= \frac{TN}{PN} = 1 - FOR$	Markedness (MK), deltaP (Δp) $= PPV + NPV - 1$	Diagnostic odds ratio (DOR) $= \frac{LR+}{LR-}$
	Balanced accuracy (BA) $= \frac{TPR + TNR}{2}$	F_1 score $= \frac{2 PPV \times TPR}{PPV + TPR} = \frac{2 TP}{2 TP + FP + FN}$	Fowlkes–Mallows index (FM) $= \sqrt{PPV \times TPR}$	Matthews correlation coefficient (MCC) $= \sqrt{TPR \times TNR \times PPV \times NPV} - \sqrt{FNR \times FPR \times FOR \times FDR}$	Threat score (TS), critical success index (CSI), Jaccard index $= \frac{TP}{TP + FN + FP}$

A *CoViD-19 breath analyzer* was developed. The device can analyze the exhalation of a person and makes a prediction if that person has CoViD-19 or not. For testing the device, the breaths of 10,000 persons were analyzed and additionally the persons were PCR-tested to check if they really had CoViD-19. The CoViD-19 *prevalence was 10%*. That means 1,000 from the 10,000 had CoViD-19. The following were the results in detail: 910 were breath-tested positive and had CoViD-19, 90 were breath-tested negative but had CoViD-19, 60 were breath-tested positive but were not infected with CoViD-19, and 8940 were breath-tested negative and were indeed not infected with CoViD-19.

a. Fill out the values in this *confusion matrix* with the values for TP, FP, FN, TN, P, N, PP, and PN

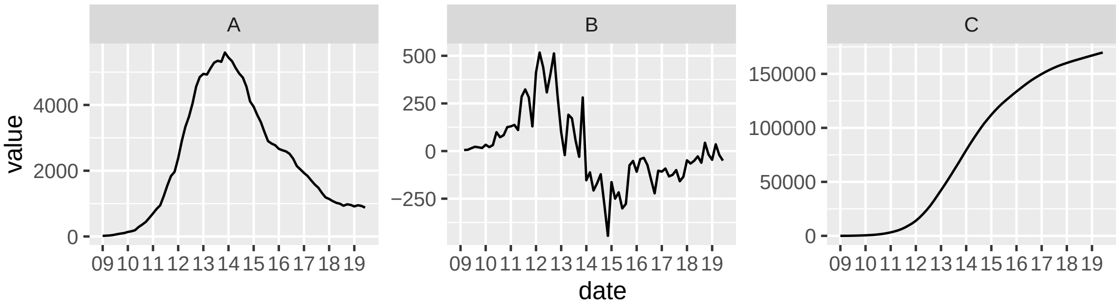
	PP =	PN =
P =		
N =		

b. Compute the *sensitivity* (fractional statmente like 100/1000 is enough)! ____________

c. Analogously, compute the *specificity*. ____________

d. Now, we assume that the computed sensitivity and the specificity are characteristics of the breath analyzer device. 10,000 new persons are tested but the the **prevalence is only 1%**. What is the expected number of true positives (TP)? ____________

9. Each facet in the following graph shows a time-series derived from the *smoothed daily new cases* of **COVID-19 in Germany** from the “Our World in Data” dataset. The data shows the first wave 2020. The x-axis marks week number of the year. For example, mark “09” marks the beginning of week 9.

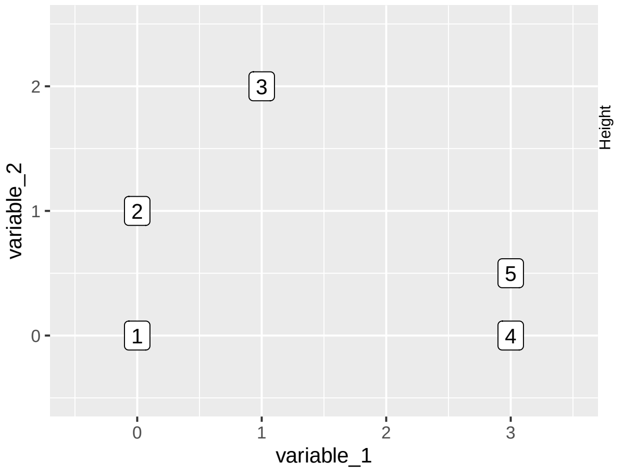


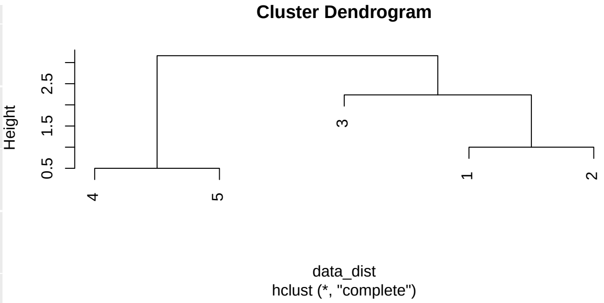
a. Time series A is called “Daily new cases”.
Give a similar name to the other two time series. The names shall be understandable for people who do not now Calculus. Descriptive name for B: ____
Descriptive name for C: ____
In which week was the maximimum for new cases: ____________________________
What happens in B when the maximum of new cases is achieved?
____

b. *Calculus question:* Which time series is the derivative of which? Which is a second derivative of which? Complete three sentences and completely describe the calculus relations:
_____ is the derivative of _____
_____ is the derivative of _____
_____ is the second derivative of _____

c. In the beginning the growth of cases is very close to exponential growth. Approximately in which week does the exponential growth ends? ______________

10. **(ANSWER ON BACKSIDE/EXTRA-SHEET)** Below you see the plot of an artificial dataset of five observations and two variables visualized in the two-dimensional coordinate system. There is also the dendrogram of *hierarchical clustering*.





Explain how hierarchical clustering works using the example data. Explain what the algorithm does step by step! Read from the dendrogram: When we want two clusters, which point(s) belong to one cluster and which to the other cluster?