2024 02 21 VB-STA5 Reexam in Statistics

Wednesday 21st of Fabruary.

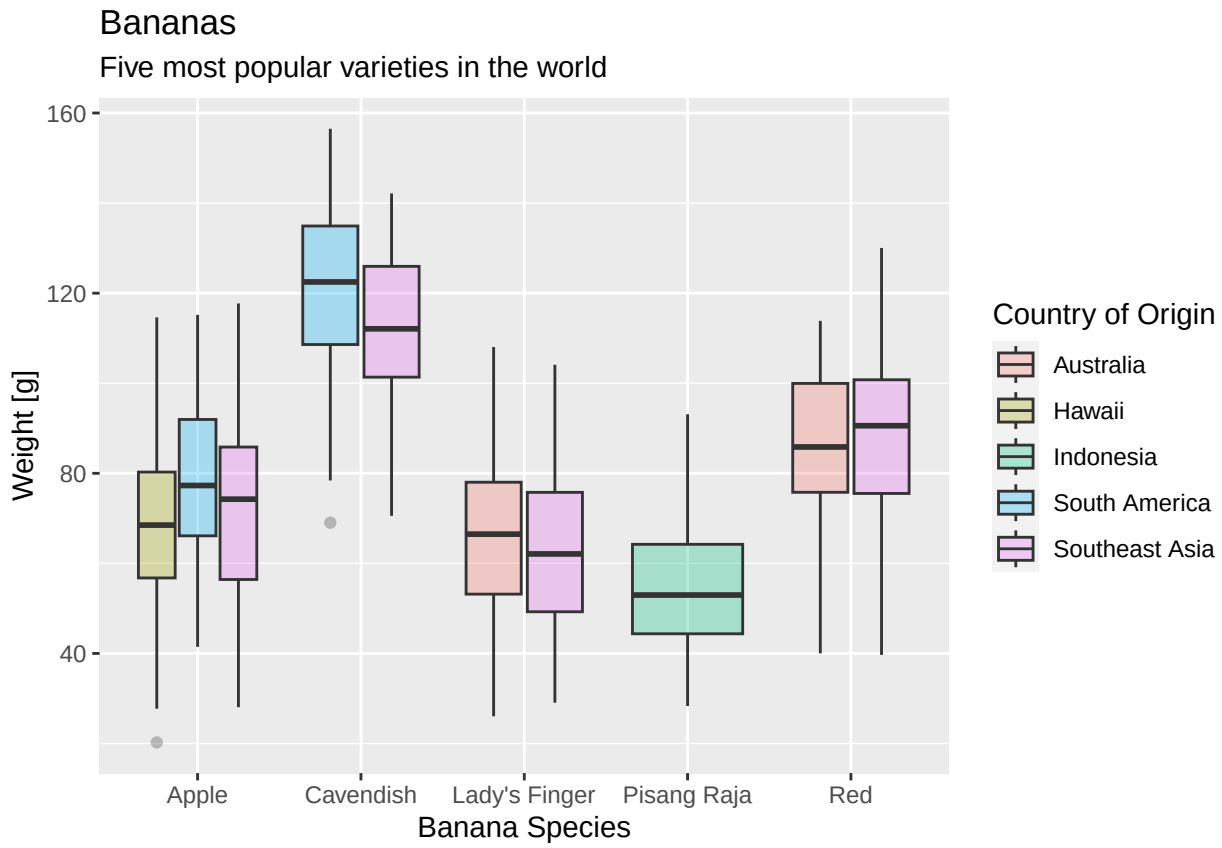
The exam set consists of 3 main exercises with 9 sub exercises in total.

Each sub exercise is weighted equally when grading the hand ins.

1. Bananas

Dataset *data/bananas_dataset.csv* contains information about five most popular banana varieties in the world.

a) Recreate the plot:



b) Describe the plot. Include description of what is presented, how it is presented, is there grouping, and what information about those groups can be read.

c) For bananas grown in Southeast Asia, present the average length and average weight divided according to Species.

The example shows similar table, but for 'Australia'

Species	Average Length [cm]	Average Weight [g]
Lady's Finger	11.76318	65.93221
Red	14.66845	86.29905

d) Banana plant originated in Southeast Asia, however Ecuador is the biggest producer of bananas in the world. Select a relevant statistical test and use it to test whether average Ecuadorian (South American) *Cavendish* bananas are bigger/heavier than average Southeast Asia *Cavendish* bananas. Form hypothesis, check for conditions, conduct a statistical test, and form conclusions.

2. USA population in 2020

The dataset *data/US_state_capitol_population_2020.csv* contains information about population and race in state capitols in US. Population of entire USA in 2020 is summarized in a table below:

Race	USA
White	195223627
Black or African American	45077102
American Indian and Alaska Native	4308841
Asian	20881305
Native Hawaiian and Other Pacific Islander	994348
Two or More Races	9943478
Hispanic or Latino	63306813

a) Select a relevant statistical test, and use it to check, whether Race distribution of Atlanta population is following the same distribution as Race distribution of the entire country. Form hypothesis, check for conditions, conduct a statistical test, and form conclusions.

3. Airfares

airq412.csv contains information about airfares and passengers for the U.S. Domestic Routes for 4th quarter of 2012. Norwegian Airlines wants to break into the U.S. market with a new route in between Point Place, Wisconsin (-) and Los Angeles, California (LAX). Currently there are **no commercial** flights from Point Place, and due to that the city is not included in the database.

The distance in between two cities is 2260 miles and is expected to have approximately 150 passengers per week.

a) Create a linear model to predict 'Average Fare' using 'Distance'. Then evaluate the model (check if conditions are fulfilled).

b) Evaluate the model from exercise 3a (check if conditions are fulfilled).

c) Propose a price for a ticket from Point Place to Los Angeles using the model created in exercise 3a.

d) Create two multiple regression models to predict *Average Fare* using:

- *Distance*, *Average Weekly Passengers*, *Market Share MLA*,
- *Distance*, *Market Share MLA*

Which one is better in your opinion and why? Use chosen model to predict an average fare for new route between Point Place and Los Angeles.