2022 02 23 VB-STA5 Reexam in Statistics

Wednesday 23rd of February.

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. Mind the gap.

Dataset *data/London_transport_passengers.csv* contains information about journeys taken using London Transport since 2010. Dataset *data/London_transport_codes.csv* contains information about codes for London Transport types.

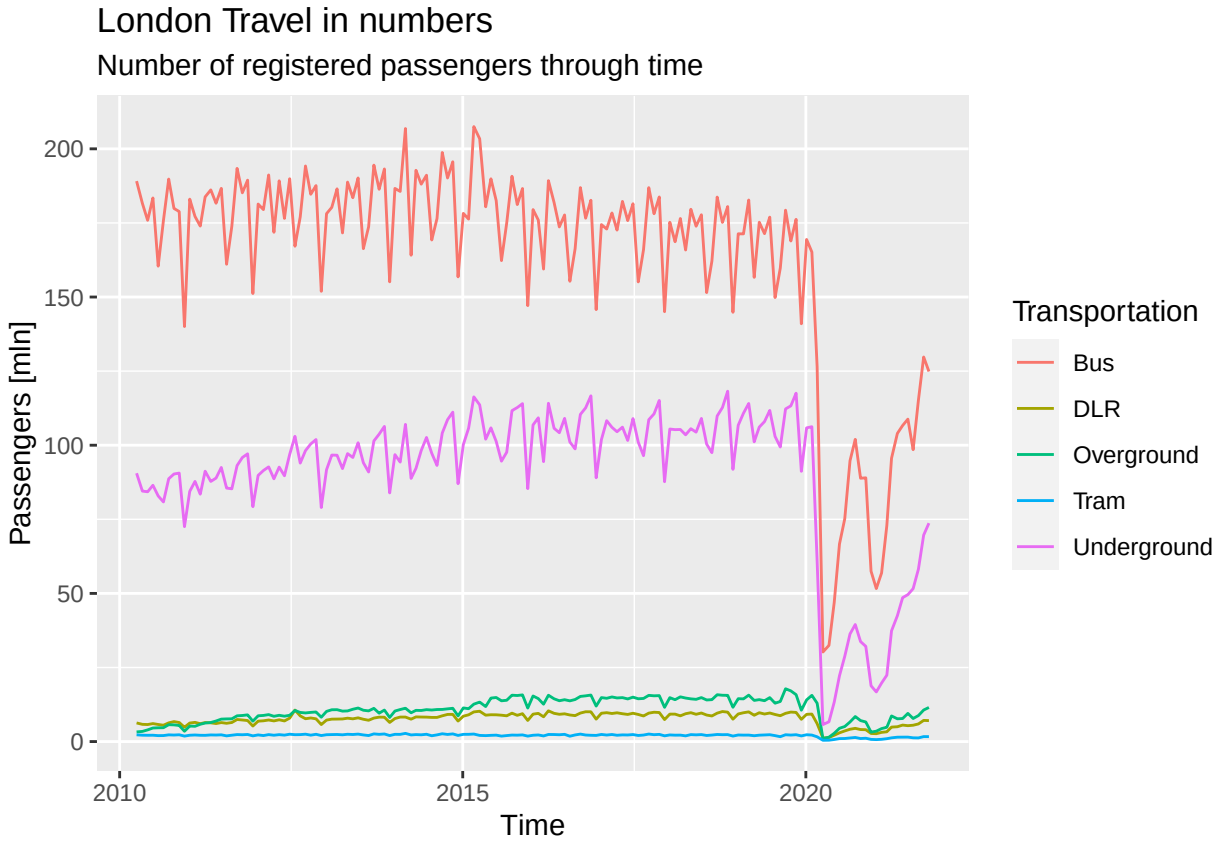
a) Join the two datasets.

b) Present average number of passengers on all modes of transport in Reporting period *11* through the years in descending order.

The example shows the mean number of passengers in period 9 in ascending order.

Transportation	Mean number of passengers [mln]
Tram	2.266618
DLR	8.336597
Overground	11.936822
Underground	101.886974
Bus	176.683535

c) Recreate the plot.



d) Describe the plot.

e) In 2017 a group of students at Imperial College London conducted a survey of 927 London commuters. The survey asked questions about service satisfaction using a couple of carefully formed questions. The students asked random travelers to answer the questions. Here is a summary of how many people were surveyed in the different modes of transport:

Mode of Transport	No. of passengers interviewed
Bus	461
Underground	347
DLR	26
Tram	25
Overground	68

Is this a representative sample of the population of passengers in 2017? Conduct a test to form statistical conclusions.

2. Ice cream

Dataset *data/ice_cream.csv* contains information from an experiment, where a group of people were asked to choose in between three flavours of ice cream - strawberry, chocolate, and vanilla. Subsequently, they have been evaluated in playing video games and doing puzzles.

a) Plot the density functions of video game scores for each ice cream flavour.

b) Is there a statistically significant difference between the mean puzzle score for males with vanilla preference vs. males with strawberry preference? Conduct a suitable test.

3. Health insurance

Dataset *data/insurance.csv* contains information about over 1000 randomly chosen U.S. policyholders. Their insurance packages range in between low-cost insurance - up to $15.000 per year, medium-cost insurance $15.000 - $30.000 per year, and high-cost insurance - above $30.000 per year.

a) What are statistically significant predictors of the high-cost insurance? Create a model and tune it.

b) Evaluate the model.