Exercises Day 3*

Ronald Geskus Tran Thai Hung

Last version: April 04, 2025

Table of contents

Instructions

Import the data set and load the **ggplot2** package. Change the path to the Excel file if it is in another folder on your computer.

```
#| echo: true
library(readxl)
library(ggplot2)
titanic <- read_excel("data/raw/Titanic3.xlsx", na="NA")
```

*Please do not circulate.

Exercise 1

Plot the price that was paid (variable `fare`) as a function of age. Assign/map the appropriate variable to the x-axis and y-axis. Next, define the geometric object. Change the label of fare into "fare paid (GBP)". In R:

```
ggplot(...) + geom_point(...) + ylab(...)
```

See what happens if you change the coordinate system to polar via `+ coord_polar()`.

Exercise 2

Return to the cartesian coordinate system. Add a second layer that fits the linear regression line with 95% confidence intervals through the scatterplot (this is seen as smoothing, choose as method `lm`). Change the overall theme, trying `theme_bw` and `theme_minimal`.

In R: add `... + geom_smooth(method="lm") + ...`. Save the graph as a png file using the `ggsave` function. .

Exercise 3

The distribution of fare is very skewed. Therefore, it may be better to plot fare on a logarithmic scale. Repeat the exercise but now use a log scale along the fare axis. If you compare the fitted linear trends, you will notice that they do not coincide. This is for two reasons. First, there were persons that didn't pay any money. They are removed if the logarithm is taken (you cannot take the log of zero). Second, the regression line is computed after the scale transformation, hence based on very different values.

In R: replace `ylab(...)` by `scale_y_log10(...)`

Exercise 4

We continue with the graph from the last exercise that uses the log scale for fare. Color the individual data by passenger class. Split the calculation and plot of the linear regression line by these three groups.

In R: add `color=pclass` to the aesthetic attributes.

Exercise 5

Split the graph into two separate ones by sex, next to each other. Add a rug plot to both sides.

In R: add `... + facet_wrap(~sex) + geom_rug()`

Exercise 6

Make a boxplot of age (along y-axis) for each of the passenger classes (along x-axis). See what happens if you flip the coordinates via `coord_flip()`.

Exercise 7

Add the individual data points to the boxplot, using `geom_point(...)`. What do you observe? Change the default theme if you think it appropriate.

Use the jitter instead of the point geometric object to be able to discriminate them, give the points a red color and make them transparent by setting the transparency parameter (alpha) to 0.2.

Replace the boxplot by a violin plot.

Exercise 8

We make a new graph in which we perform an exploratory analysis of the relation between survival and age, sex and passenger class. Define the assignment/mapping to the x-axis (`age`) and y-axis (`survived`). Make a scatterplot of survival status as a function of age.

In order to better see the individual outcomes, use jitter with a height parameter of 0.05 and a width parameter of 0 (why would you set width at 0?); set transparency alpha to 0.5. Add a smooth layer using the **loess** method with 95% confidence intervals.

Exercise 9

Create separate loess curves for males and females. Put the legend above the graph using `theme(legend.position='top')`.

Exercise 10

Split the graph into three panels based on passenger class. Truncate the range of the y-axis. Do this by setting the y-axis limits to c(-0.1,1.1) and see what happens.