**Titanic Survival Exercises**

Put all your new skills together to perform exploratory data analysis on a classic machine learning dataset: Titanic survival!

**Background**

The Titanic was a British ocean liner that struck an iceberg and sunk on its maiden voyage in 1912 from the United Kingdom to New York. More than 1,500 of the estimated 2,224 passengers and crew died in the accident, making this one of the largest maritime disasters ever outside of war. The ship carried a wide range of passengers of all ages and both genders, from luxury travelers in first-class to immigrants in the lower classes. However, not all passengers were equally likely to survive the accident. We use real data about a selection of 891 passengers to learn who was on the Titanic and which passengers were more likely to survive.

**Libraries, Options, and Data**

Define the titanic dataset starting from the **titanic** library with the following code:

options(digits = 3) # report 3 significant digits  
library(tidyverse)  
library(titanic)

titanic <- titanic\_train %>%  
    select(Survived, Pclass, Sex, Age, SibSp, Parch, Fare) %>%  
    mutate(Survived = factor(Survived),  
           Pclass = factor(Pclass),  
           Sex = factor(Sex))

correct

Pclass

incorrect

Sex

correct

SibSp

correct

Parch

correct

Fare

correct

You have used 3 of 3 attempts Some problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

**Question 2: Demographics of Titanic Passengers**

3.01/3.5 points (graded)

Make density plots of age grouped by sex. Try experimenting with combinations of faceting, alpha blending, stacking and using variable counts on the y-axis to answer the following questions. Some questions may be easier to answer with different versions of the density plot.

Which of the following are true?

Select all correct answers.

Females and males had the same general shape of age distribution.

The age distribution was bimodal, with one mode around 25 years of age and a second smaller mode around 5 years of age.

There were more females than males.

The count of males of age 40 was higher than the count of females of age 40.

The proportion of males age 18-35 was higher than the proportion of females age 18-35.

The proportion of females under age 17 was higher than the proportion of males under age 17.

The oldest passengers were female.

partially correct

Answer

Incorrect:

Try again. Try putting count on the y-axis instead of density.

You have used 3 of 3 attempts Some problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

**Question 3: QQ-plot of Age Distribution**

1/1 point (graded)

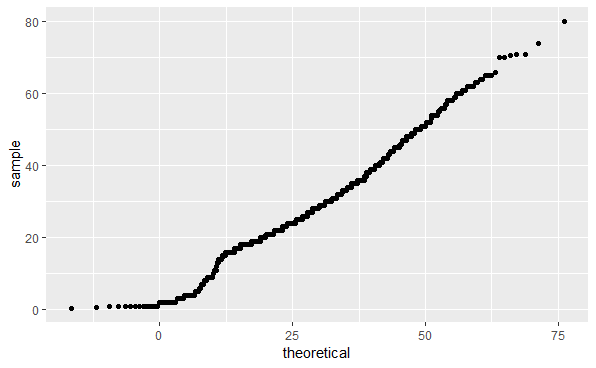
Use geom\_qq to make a QQ-plot of passenger age and add an identity line with geom\_abline. Filter out any individuals with an age of NA first. Use the following object as the dparams argument in geom\_qq:

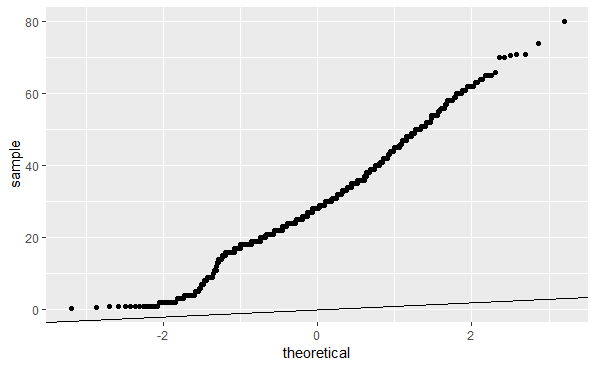
params <- titanic %>%

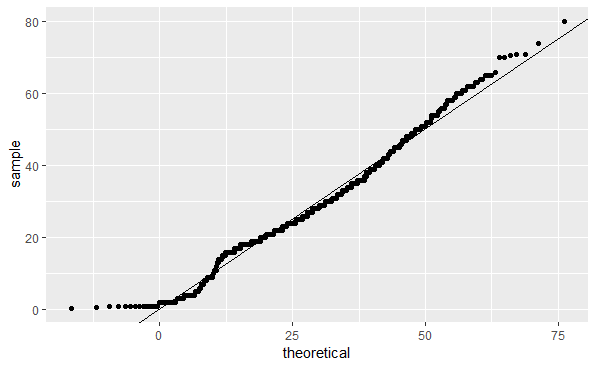
filter(!is.na(Age)) %>%

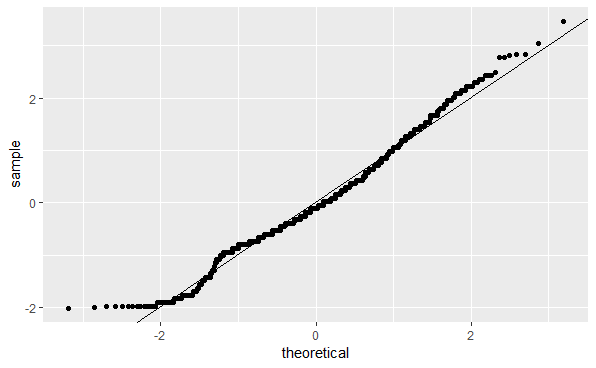
summarize(mean = mean(Age), sd = sd(Age))

Which of the following is the correct plot according to the instructions above?









unanswered

You have used 1 of 2 attempts Some problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Your answers were previously saved. Click 'Submit' to grade them.

**Question 4: Survival by Sex**

2.0/2.0 points (graded)

To answer the following questions, make barplots of the Survived and Sex variables using geom\_bar. Try plotting one variable and filling by the other variable. You may want to try the default plot, then try adding position = position\_dodge() to geom\_bar to make separate bars for each group.

You can read more about making barplots in the [textbook section on ggplot2 geometries](https://rafalab.github.io/dsbook/distributions.html#other-geometries).

Which of the following are true?

Select all correct answers.

Less than half of passengers survived.

Most of the survivors were female.

Most of the males survived.

Most of the females survived.

correct

You have used 1 of 2 attempts Some problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

**Question 5: Survival by Age**

3.0/3.0 points (graded)

Make a density plot of age filled by survival status. Change the y-axis to count and set alpha = 0.2.

Which age group is the only group more likely to survive than die?

0-8

10-18

18-30

30-50

50-70

70-80

correct

Which age group had the most deaths?

0-8

10-18

18-30

30-50

50-70

70-80

correct

Which age group had the highest proportion of deaths?

0-8

10-18

18-30

30-50

50-70

70-80

correct

You have used 2 of 4 attempts Some problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

**Question 6: Survival by Fare**

1.5/2.5 points (graded)

Filter the data to remove individuals who paid a fare of 0. Make a boxplot of fare grouped by survival status. Try a log2 transformation of fares. Add the data points with jitter and alpha blending.

Which of the following are true?

Select all correct answers.

Passengers who survived generally payed higher fares than those who did not survive.

The interquartile range for fares was smaller for passengers who survived.

The median fare was lower for passengers who did not survive.

Only one individual paid a fare around $500. That individual survived.

Most individuals who paid a fare around $8 did not survive.

partially correct

Answer

Incorrect:

Try again. The interquartile range (IQR) is the size of the box.

Try again. Make sure you add points with geom\_jitter.

You have used 2 of 2 attempts Some problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

**Question 7: Survival by Passenger Class**

2.0100000000000002/3.0 points (graded)

The Pclass variable corresponds to the passenger class. Make three barplots. For the first, make a basic barplot of passenger class filled by survival. For the second, make the same barplot but use the argument position = position\_fill() to show relative proportions in each group instead of counts. For the third, make a barplot of survival filled by passenger class using position = position\_fill().

You can read more about making barplots in the [textbook section on ggplot2 geometries](https://rafalab.github.io/dsbook/distributions.html#other-geometries).

Which of the following are true?

Select all correct answers.

There were more third class passengers than passengers in the first two classes combined.

There were the fewest passengers in first class, second-most passengers in second class, and most passengers in third class.

Survival proportion was highest for first class passengers, followed by second class. Third-class had the lowest survival proportion.

Most passengers in first class survived. Most passengers in other classes did not survive.

The majority of survivors were from first class. (Majority means over 50%.)

The majority of those who did not survive were from third class.

partially correct

Answer

Incorrect:

Try again. Look at the first barplot of passenger class filled by survival without using position\_fill.

Try again. Look at the third barplot of survival filled by passenger class using position\_fill to plot proportions rather than counts.

You have used 2 of 2 attempts Some problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

**Question 8: Survival by Age, Sex and Passenger Class**

2.5/2.5 points (graded)

Create a grid of density plots for age, filled by survival status, with count on the y-axis, faceted by sex and passenger class.

Which of the following are true?

Select all correct answers.

The largest group of passengers was third-class males.

The age distribution is the same across passenger classes.

The gender distribution is the same across passenger classes.

Most first-class and second-class females survived.

Almost all second-class males did not survive, with the exception of children.

unanswered

You have used 1 of 2 attempts