Assessment due Jun 5, 2021 13:54 +03

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

Be sure that you have installed the **titanic** package before proceeding.

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))
```

Question 1: Variable Types

3/3 points (graded)

Inspect the data and also use `?titanic_train` to learn more about the variables in the dataset. Match these variables from the dataset to their variable type. There is at least one variable of each type (ordinal categorical, non-ordinal categorical, continuous, discrete).

Survived

non-ordinal categorical ✔ **Answer:** non-ordinal categorical

Pclass

ordinal categorical ✔ **Answer:** ordinal categorical

Sex

non-ordinal categorical ✔ **Answer:** non-ordinal categorical

SibSp

discrete ✔ **Answer:** discrete

Parch

discrete ✔ **Answer:** discrete

Fare

continuous ✔ **Answer:** continuous

Submit You have used 1 of 3 attempts

Answers are displayed within the problem

Question 2: Demographics of Titanic Passengers

3.5/3.5 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.

- [x] Females and males had the same general shape of age distribution.
- [x] 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.
- [x] The count of males of age 40 was higher than the count of females of age 40.
- [x] The proportion of males age 18-35 was higher than the proportion of females age 18-35.
- [x] The proportion of females under age 17 was higher than the proportion of males under age 17.
- [ ] The oldest passengers were female.

✔

Answer Code and Explanation

A faceted plot is useful for comparing the distributions of males and females for A. Each sex has the same general shape with two modes at the same locations, though proportions differ slightly across ages and there are more males than females.

```
        titanic %>%
ggplot(aes(Age, fill = Sex)) +
geom_density(alpha = 0.2) +
facet_grid(Sex ~ .)
```

A stacked density plot with count on the y-axis is useful for answering B, C and D. The main mode is around age 25 and a second smaller mode is around age 4-5. There are more males than females as indicated by a higher total area and higher counts at almost all ages. With count on the y-axis, it is clear that more males than females are age 40.

```
        titanic %>%
ggplot(aes(Age, y = ..count.., fill = Sex)) +
geom_density(alpha = 0.2, position = "stack")
```

A plot filled by sex with alpha blending helps reveal the answers to E, F and G. There is a higher proportion of females than males below age 17, a higher proportion of males than females for ages 18-35, approximately the same proportion of males and females age 35-55, and a higher proportion of males over age 55. The oldest individuals are male.

```
        titanic %>%
ggplot(aes(Age, fill = Sex)) +
geom_density(alpha = 0.2)
```

Submit

You have used 2 of 3 attempts

Answers are displayed within the problem

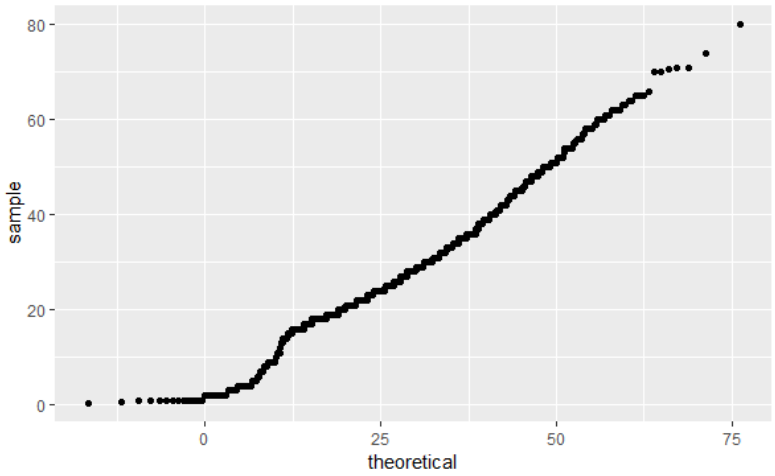
Question 3: QQ-plot of Age Distribution

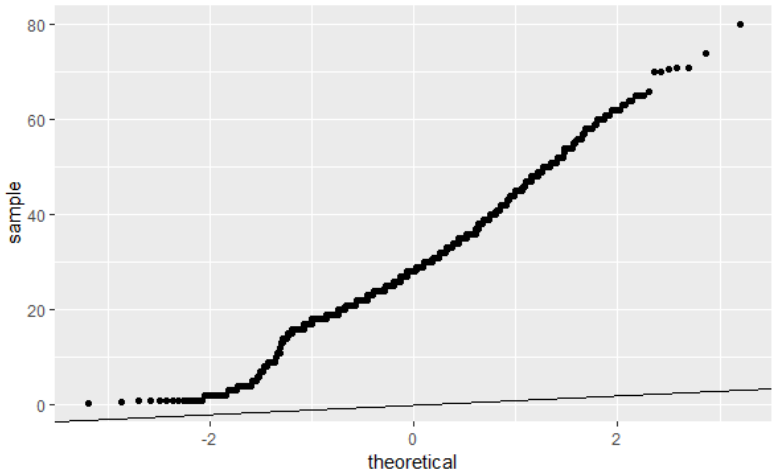
1/1 point (graded)

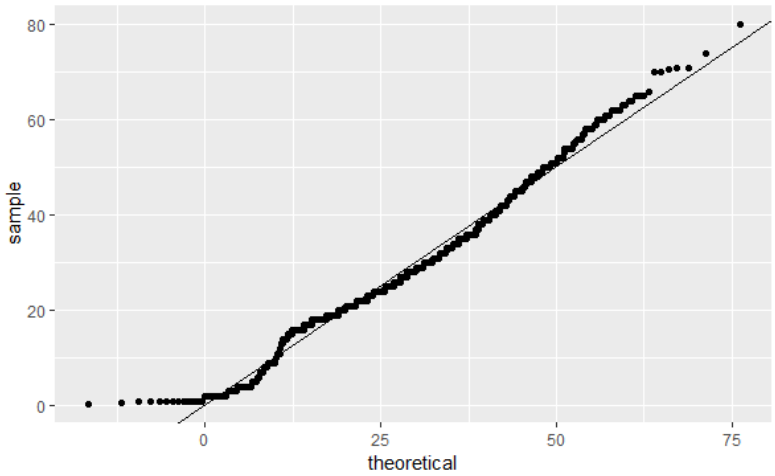
Use `geom_qq()` to make a QQ-plot of passenger age and add an identity line with `geom_abline()`. Filter out any individuals with an age of NA first. Use the following object as the `dparams` argument in `geom_qq()`:

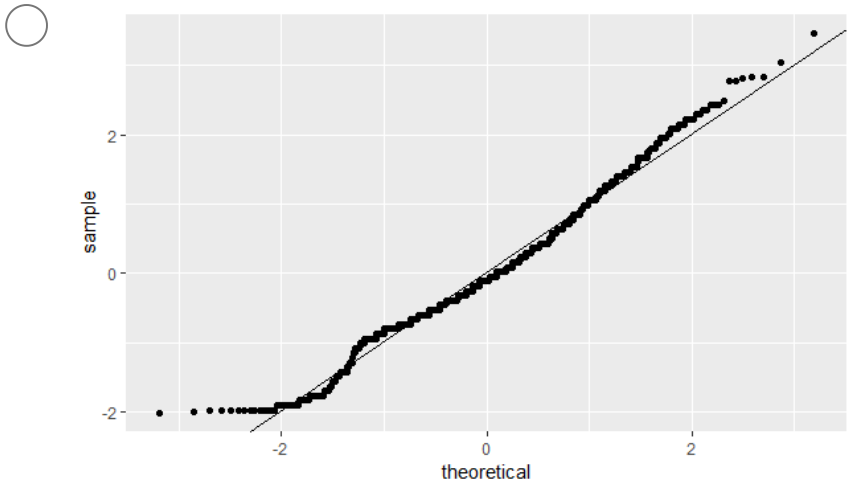
```
   params <- titanic %>%
filter(!is.na(Age)) %>%
summarize(mean = mean(Age), sd = sd(Age))
```

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


sample
theoretical
0
20
40
60
80
0
25
50
75


sample
theoretical
0
20
40
60
80
-2
0
2


sample
theoretical
0
20
40
60
80
0
25
50
75



✔

Answer Code

```
        titanic %>%
  filter(!is.na(Age)) %>%
  ggplot(aes(sample = Age)) +
  geom_qq(dparams = params) +
  geom_abline()
```

Submit You have used 1 of 2 attempts

Answers are displayed within the problem

Question 4: Survival by Sex

2/2 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.

Which of the following are true?

Which of the following are true?

Select all correct answers.

- [x] Less than half of passengers survived.
- [x] Most of the survivors were female.
- [ ] Most of the males survived.
- [x] Most of the females survived.

✔

Explanation and Answer Code

A and B can be clearly seen in the barplot of survival status filled by sex. The count of survivors is lower than the count of non-survivors. The bar of survivors is more than half filled by females. Alternatively, the bars can be split by sex with `position_dodge`, showing the "Female, Survived" bar has a greater height than the "Male, survived" bar. C and D are more clearly seen in the barplot of sex filled by survival status, though they can also be determined from the first barplot. Most males did not survive, but most females did survive.

```
            #plot 1 - survival filled by sex
titanic %>%
    ggplot(aes(Survived, fill = Sex)) +
    geom_bar()
# plot 2 - survival filled by sex with position_dodge
titanic %>%
    ggplot(aes(Survived, fill = Sex)) +
    geom_bar(position = position_dodge())
#plot 3 - sex filled by survival
titanic %>%
    ggplot(aes(Sex, fill = Survived)) +
    geom_bar()
```

Submit

You have used 1 of 2 attempts

Answers are displayed within the problem

Question 5: Survival by Age

3/3 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

✔

Which age group had the most deaths?

- ( ) 0-8
- ( ) 10-18
- (•) 18-30
- ( ) 30-50
- ( ) 50-70
- ( ) 70-80

✔

Which age group had the highest proportion of deaths?

- ○ 0-8
- ○ 10-18
- ○ 18-30
- ○ 30-50
- ○ 50-70
- ◉ 70-80

✔

Answer Code

```
        titanic %>%
ggplot(aes(Age, y = ..count.., fill = Survived)) +
geom_density(alpha = 0.2)
```

Submit

You have used 1 of 4 attempts

ℹ Answers are displayed within the problem

Question 6: Survival by Fare

2.5/2.5 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.

- [x] Passengers who survived generally payed higher fares than those who did not survive.
- [ ] The interquartile range for fares was smaller for passengers who survived.
- [x] The median fare was lower for passengers who did not survive.
- [ ] Only one individual paid a fare around $500. That individual survived.
- [x] Most individuals who paid a fare around $8 did not survive.

✔

Answer Code

```
        titanic %>%
  filter(Fare > 0) %>%
  ggplot(aes(Survived, Fare)) +
  geom_boxplot() +
  scale_y_continuous(trans = "log2") +
  geom_jitter(alpha = 0.2)
```

Submit

You have used 1 of 2 attempts

ℹ Answers are displayed within the problem

Question 7: Survival by Passenger Class

3/3 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.

Which of the following are true?
Select all correct answers.

- [x] 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.
- [x] Survival proportion was highest for first class passengers, followed by second class. Third-class had the lowest survival proportion.
- [x] 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%.)
- [x] The majority of those who did not survive were from third class.

✔

Answer Code

```
            # barplot of passenger class filled by survival
titanic %>%
    ggplot(aes(Pclass, fill = Survived)) +
    geom_bar() +
    ylab("Proportion")
# barplot of passenger class filled by survival with position_fill
titanic %>%
    ggplot(aes(Pclass, fill = Survived)) +
    geom_bar(position = position_fill()) +
    ylab("Proportion")
# Barplot of survival filled by passenger class with position_fill
titanic %>%
    ggplot(aes(Survived, fill = Pclass)) +
    geom_bar(position = position_fill()) +
    ylab("Proportion")
```

Submit You have used 1 of 2 attempts

Answers are displayed within the problem

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 th axis, faceted by sex and passenger class.

Which of the following are true?

Select all correct answers.

- [x] 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.
- [x] Most first-class and second-class females survived.
- [x] Almost all second-class males did not survive, with the exception of children.

Answer Code

```
        titanic %>%
  ggplot(aes(Age, y = ..count.., fill = Survived)) +
  geom_density(position = "stack") +
  facet_grid(Sex ~ Pclass)
```

Submit You have used 1 of 2 attempts

- Answers are displayed within the problem