Assignment 4 - Module 4

Kazeem Abiodun Folarin

5/8/2021

Business Problem

Context

Kazo Manufacturing, a Maritime Construction company, is attempting to bring the world's first all-electric cruiseliner to the market. Instead of using a conventional diesel engine to power its twin propellers, the company is considering the prospect of using Lithium-Ion batteries and solar panels to serve as its main source of power. Its design will be catered to families that want to spend some time with their children on vacation during peak seasons on the East Coast. The Environment, Health & Safety (EHS) team needs to design its layout to accommodate the safe departure of its crew and members in the event of an imminent wreckage

Key questions

The Lead Data Scientist has tasked you producing insights to the following questions,

1. Are families with more children at a higher risk of perishing?
2. What factors of cabin members can be trusted to influence the direction of their survivability?

Choice of data set

After comparing the current schematic of the ship with data sets of known emergencies involving cruiseliners at sea, you conclude that the accommodation layout and ship design closely resembles that of the titanic. However, modern regulation requires all cabins, regardless of status, to be an equal distance away from lifeboats. After much research, you settle on a data-set of cabin members that survived the wreckage with features that could help provide the kind of insights the lead requires.

```
# dependencies
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 4.0.5
```

Data Ingestion

Summary

The data set contains 887 observations of guests onboard the Titanic. It presents features that describe their cabin class, name, sex, age, whether they had siblings or parents, how much their fare was, and whether they survived the wreckage.

```
# data paths
path_titanic_data <- "C:/Users/pc/Downloads/raw_titanic.csv"

# read titanic data-set
df_raw_titanic <- read.csv(path_titanic_data, stringsAsFactors = FALSE)
summary(df_raw_titanic)
```

```
##     Survived          Pclass          Name               Sex
##  Min.   :0.0000   Min.   :1.000   Length:887         Length:887
##  1st Qu.:0.0000   1st Qu.:2.000   Class :character   Class :character
##  Median :0.0000   Median :3.000   Mode  :character   Mode  :character
##  Mean   :0.3856   Mean   :2.306
##  3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :1.0000   Max.   :3.000
##       Age        Siblings.Spouses.Aboard Parents.Children.Aboard
##  Min.   : 0.42   Min.   :0.0000          Min.   :0.0000
##  1st Qu.:20.25   1st Qu.:0.0000          1st Qu.:0.0000
##  Median :28.00   Median :0.0000          Median :0.0000
##  Mean   :29.47   Mean   :0.5254          Mean   :0.3833
##  3rd Qu.:38.00   3rd Qu.:1.0000          3rd Qu.:0.0000
##  Max.   :80.00   Max.   :8.0000          Max.   :6.0000
##       Fare
##  Min.   :  0.000
##  1st Qu.:  7.925
##  Median : 14.454
##  Mean   : 32.305
##  3rd Qu.: 31.137
##  Max.   :512.329
```

Data Preperation

Feature selection

The following features are to be excluded from the data set,

1. Pclass - Access to critical safety systems is the same for all cabin and crew, regardless of whether they're on the high or low end of luxury.
2. Name - We are not required to produce recommendations based on unstructured data.

Data cleaning

Excluding these features produces the data set to be used for further statistical analysis to extract insights that align with key questions.

```
# rename columns
names(df_raw_titanic)[names(df_raw_titanic) == "Siblings/Spouses Aboard"] <- "Siblings_Spouses_A
board"
names(df_raw_titanic)[names(df_raw_titanic) == "Parents/Children  Aboard"] <- "Parents_Children_
Aboard"

# define features required
features_required <- c('Survived', 'Sex', 'Age', 'Siblings.Spouses.Aboard', 'Parents.Children.Ab
oard', 'Fare')
df_source <- df_raw_titanic[features_required]

# data frame prep
df <- data.frame(df_source)

# data type conversions
df$Sex <- as.factor(df$Sex)
df$Survived <- as.factor(df$Survived)

str(df)
```

```
## 'data.frame':    887 obs. of  6 variables:
##  $ Survived              : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ...
##  $ Sex                   : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Age                   : num  22 38 26 35 35 27 54 2 27 14 ...
##  $ Siblings.Spouses.Aboard: int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parents.Children.Aboard: int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Fare                  : num  7.25 71.28 7.92 53.1 8.05 ...
```

Data Analysis

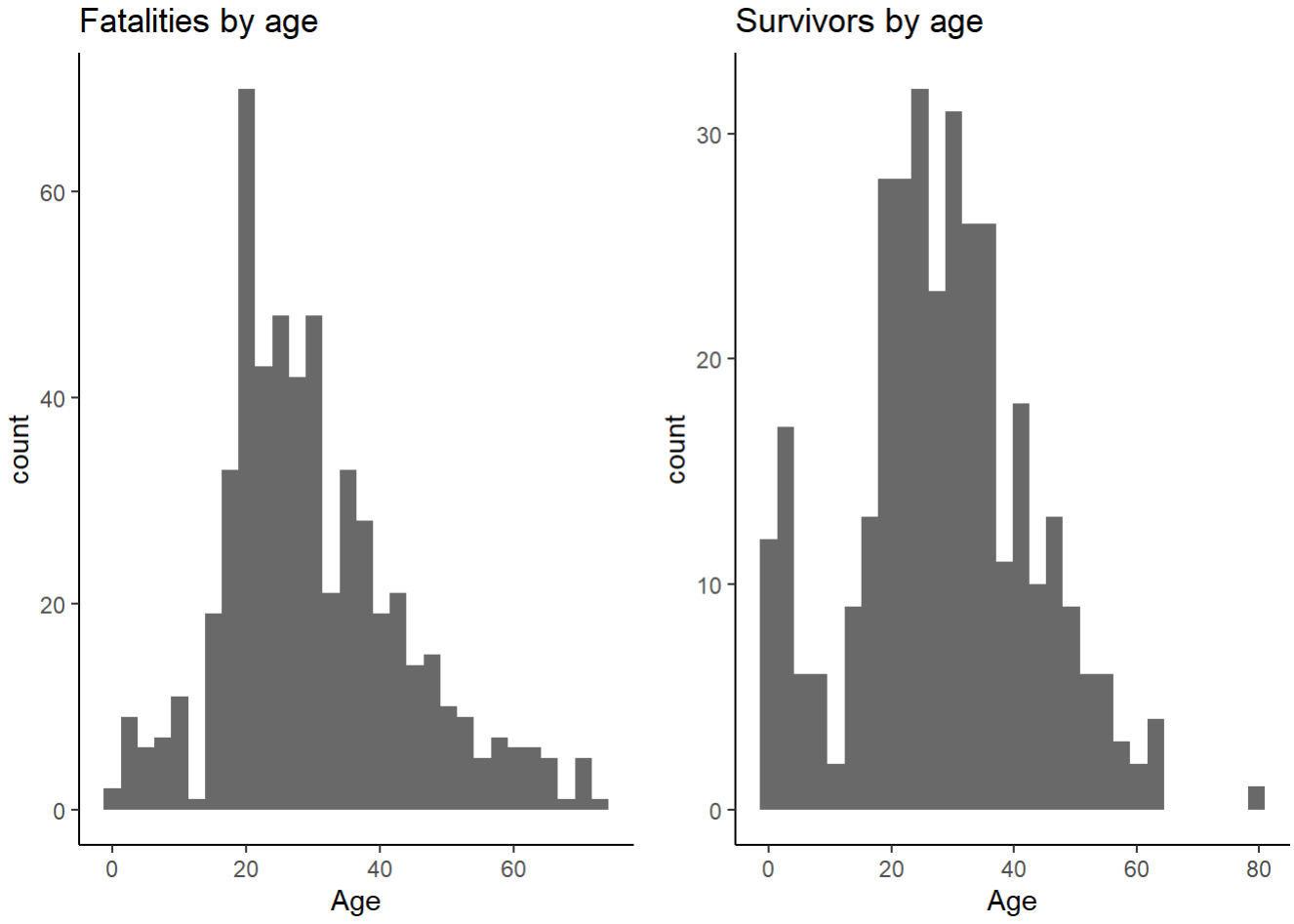
The distribution of survivability by age hints that,

1. Older adults could be at a higher risk of fatality.
2. Young adults could be prone to a lower expectation of survival.

```
age_not_survived <- ggplot(
  subset(df, Survived == 0),
  aes(x=Age)) +
  geom_histogram(position="dodge", alpha=0.9) +
  ggtitle("Fatalities by age") +
  theme_classic()

age_survived <- ggplot(
  subset(df, Survived == 1),
  aes(x=Age)) +
  geom_histogram(position="dodge", alpha=0.9) +
  ggtitle("Survivors by age") +
  theme_classic()

grid.arrange(age_not_survived, age_survived, nrow=1)
```



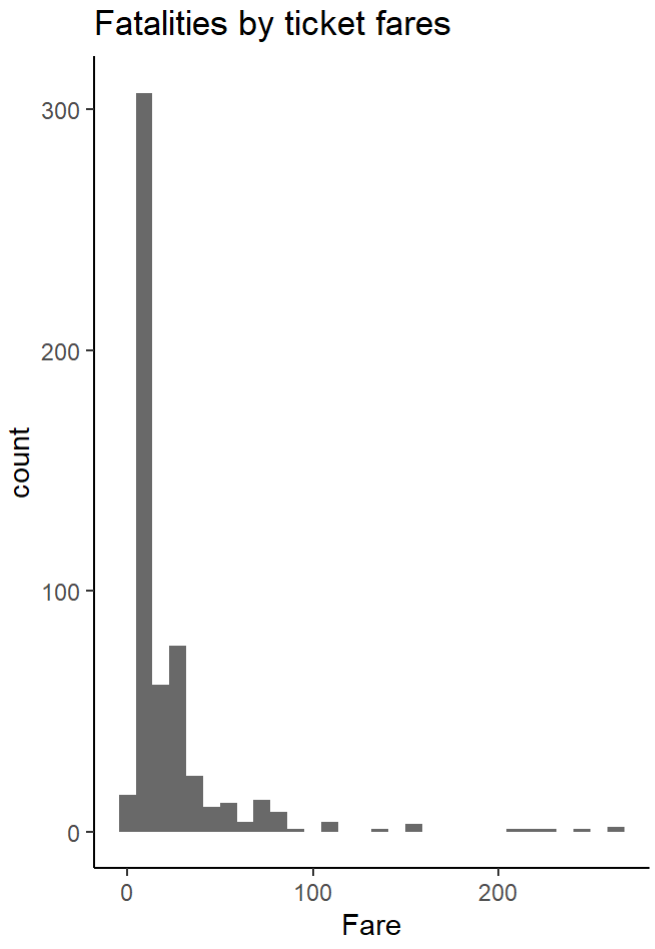
Further analysis of survivability as a function of their ticket fares suggests that,

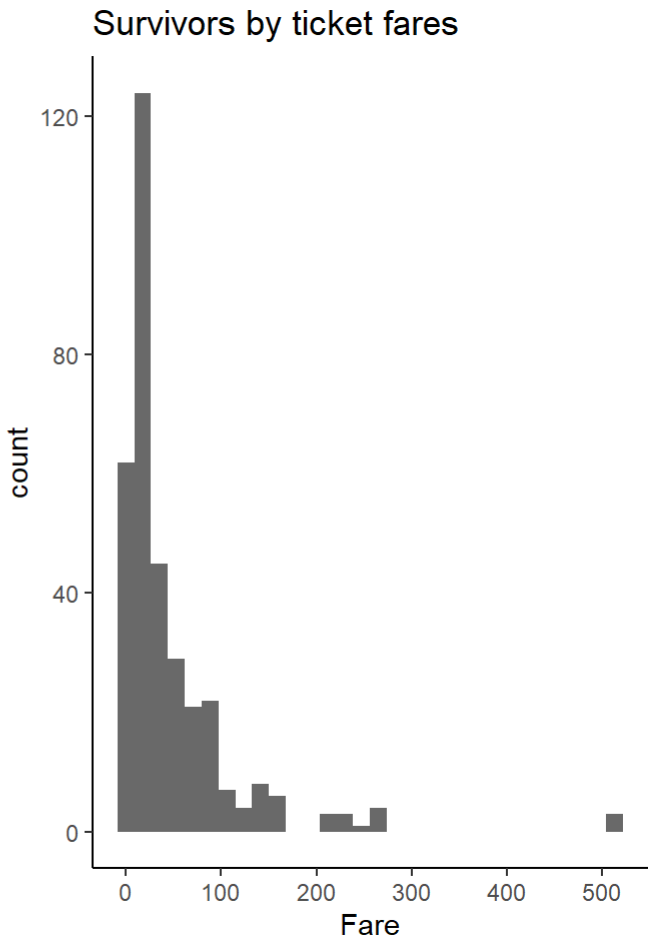
1. Those that pay significantly higher fares may be biased to an increased chance of survival.
2. Those that pay significantly lower fares could be susceptible to a higher chance of fatalities.

```
fare_not_survived <- ggplot(
  subset(df, Survived == 0),
  aes(x=Fare)) +
  geom_histogram(position="dodge", alpha=0.9) +
  ggtitle("Fatalities by ticket fares") +
  theme_classic()

fare_survived <- ggplot(
  subset(df, Survived == 1),
  aes(x=Fare)) +
  geom_histogram(position="dodge", alpha=0.9) +
  ggtitle("Survivors by ticket fares") +
  theme_classic()

grid.arrange(fare_not_survived, fare_survived, nrow=1)
```





Data Modeling

The analysis conducted above is merely descriptive, and cannot provide the kind of insights that incorporate multiple explanatory factors. Therefore, a linear regression model was developed to determine statistically significant features that impact wreckage survivability.

Note,

1. The p-values indicate the level of significance. The accepted threshold is assumed to be less than 5%.
2. The coefficients of each feature indicate the direction of survivability for either increases or decreases of the target feature, Survived.

```
# data type conversions
df$Survived <- as.integer(df$Survived)
df$Survived <- df$Survived - 1
df$Sex <- as.character(df$Sex)

# replace values and perform final conversions
df$Sex[df$Sex == "male"] <- 1
df$Sex[df$Sex == "female"] <- 0
df$Sex <- as.integer(df$Sex)

# perform and print the regression statistics
summary(lm(Survived ~ ., df))
```

```
##
## Call:
## lm(formula = Survived ~ ., data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9672 -0.2115 -0.1468  0.2582  0.9831
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              0.7951265  0.0399277  19.914  < 2e-16 ***
## Sex                     -0.5351301  0.0290160 -18.443  < 2e-16 ***
## Age                     -0.0028866  0.0010103  -2.857  0.00437 **
## Siblings.Spouses.Aboard -0.0579546  0.0137480  -4.216 2.75e-05 ***
## Parents.Children.Aboard -0.0324667  0.0188099  -1.726  0.08469 .
## Fare                     0.0019842  0.0002824   7.026 4.26e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.396 on 881 degrees of freedom
## Multiple R-squared:  0.3426, Adjusted R-squared:  0.3388
## F-statistic: 91.81 on 5 and 881 DF,  p-value: < 2.2e-16
```

Model Insights

Based on the findings of the linear regression model,

1. The sex, age, ticket fair whether a crew member has siblings or spouses on board statistically affect the survivability of the crew member.
2. Males are more likely to survive than females.
3. Younger crew members are likely to survive.
4. Members that pay a higher ticket fare are more likely to survive.

Discussion

Based on the above, the EHS team will inculcate the findings in the design to have a safe layout that accomodates; older ones by creating a Standard operating procedure (SOP) which allows crew members with aged ones or couples to have easy access to emergency exit and also making sure there are life boats at all end to accommodate all passengers. THe SOP should ensure that at no time should the number of passengers exceed the number that can accommodated on the life boats.

References

1. **Titanic data-set**, https://web.stanford.edu/class/archive/cs/cs109/cs109.1166/stuff/titanic.csv (https://web.stanford.edu/class/archive/cs/cs109/cs109.1166/stuff/titanic.csv)