

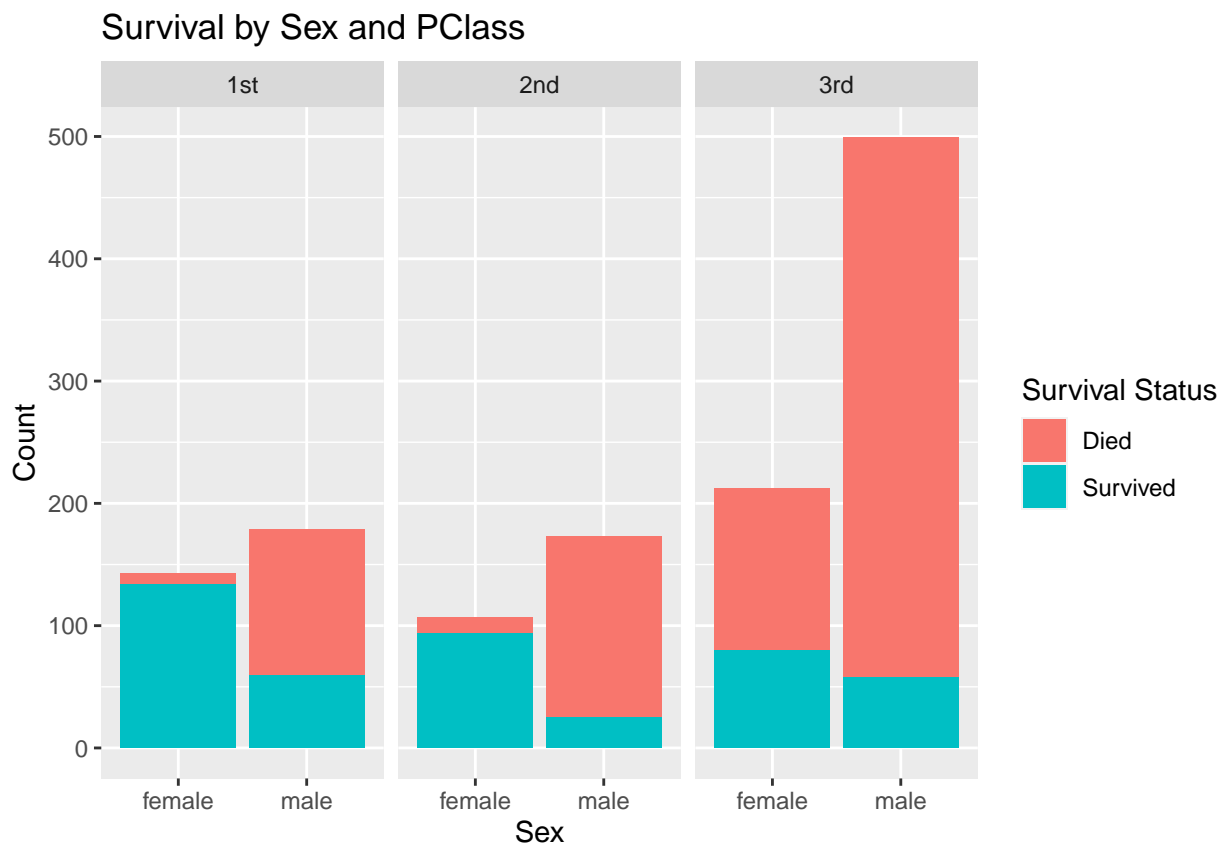
Data Analysis Report

Titanic

a) Overview + linear model for Survival

Survival based on Sex and PClass

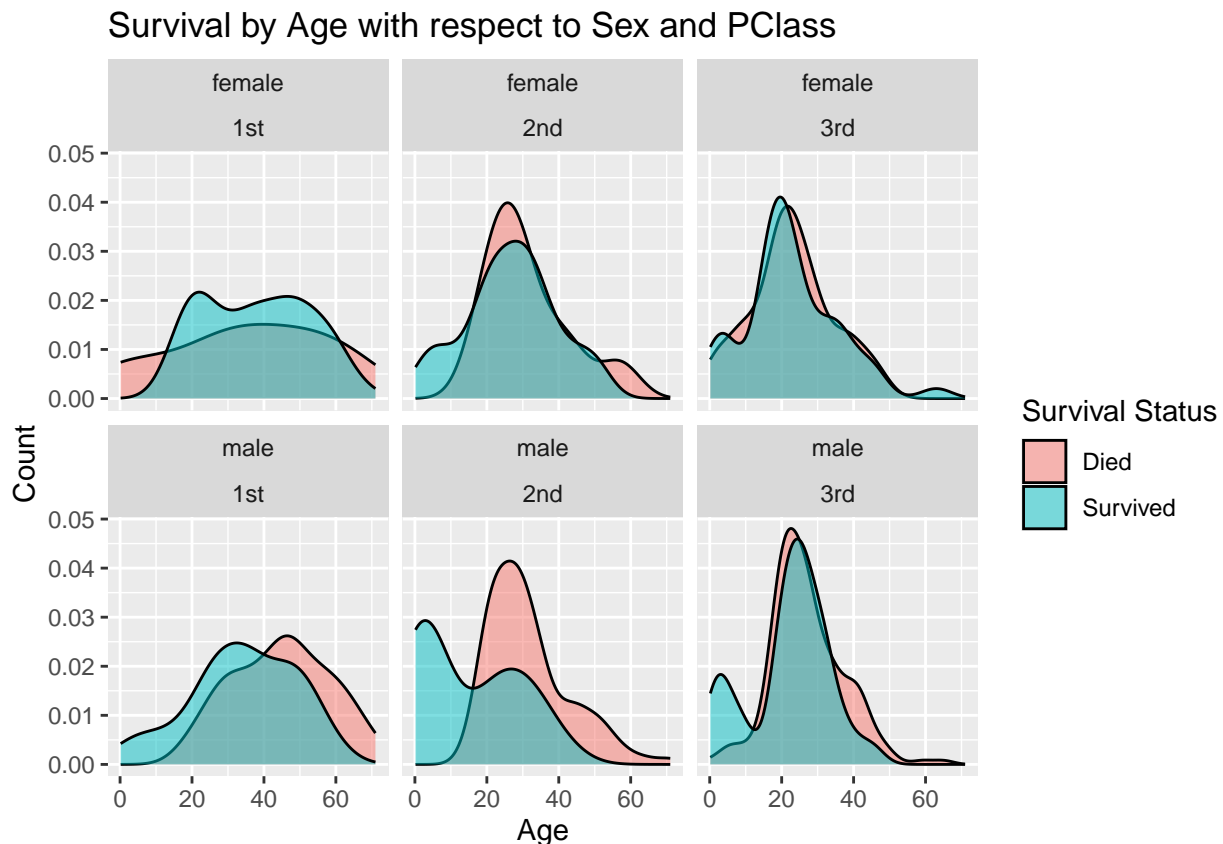
```
titanic %>%  
  mutate(Survived = factor(Survived, labels = c("Died", "Survived"))) %>%  
  ggplot(aes(x = Sex, fill = Survived)) +  
  geom_bar() +  
  facet_wrap(~PClass) +  
  labs(  
    title = "Survival by Sex and PClass",  
    x = "Sex",  
    y = "Count",  
    fill = "Survival Status"  
  )
```



Survival based on Age with respect to Sex and PClass

```
titanic %>%
  mutate(Survived = factor(Survived, labels = c("Died", "Survived"))) %>%
  ggplot(aes(x = Age, fill = Survived)) +
  geom_density(alpha = 0.5) +
  facet_wrap(~Sex + PClass) +
  labs(x = "Age",
       y = "Count",
       title = "Survival by Age with respect to Sex and PClass",
       fill = "Survival Status")
```

Warning: Removed 557 rows containing non-finite values (`stat_density()`).



We will fit a linear model, having Survived as the effect and, independently, PClass, Age and Sex as possible causes.

```
model <- glm(Survived ~ PClass + Age + Sex, data = titanic, family = binomial)
model_summary <- summary(model)$coefficients
model_summary
```

```
##           Estimate Std. Error   z value    Pr(>|z|)
## (Intercept)  3.75966210  0.397567324   9.456668 3.179129e-21
## PClass2nd    -1.29196240  0.260075781  -4.967638 6.777324e-07
## PClass3rd    -2.52141915  0.276656805  -9.113888 7.948131e-20
## Age         -0.03917681  0.007616218  -5.143868 2.691392e-07
## Sexmale      -2.63135683  0.201505379 -13.058494 5.684093e-39
```

```
age_effect_estimate <- summary(model)$coefficients["Age", "Estimate"]
age_p_value <- summary(model)$coefficients["Age", "Pr(>|z|)"]

pclass2nd_p_value <- summary(model)$coefficients["PClass2nd", "Pr(>|z|)"]
pclass3rd_p_value <- summary(model)$coefficients["PClass3rd", "Pr(>|z|)"]
sexmale_p_value <- summary(model)$coefficients["Sexmale", "Pr(>|z|)"]
```

We observe a low p_value (2.6913921×10^{-7}) for Age, therefore we reject the initial hypothesis and conclude that age has an effect on the Survival. Apparently, the chances of survival change with (-0.0391768) with each year.

We observe a low p_values for PClass2nd (6.7773237×10^{-7}), PClass3rd ($7.9481311 \times 10^{-20}$), Sexmale ($5.6840932 \times 10^{-39}$), therefore we reject the initial hypothesis and conclude that being 2nd class, 3rd class or being a male has an effect on the Survival.

b) Interactions + Studying a 55 years old person chances to survive

Interactions

```
model_age_pclass <- glm(Survived ~ Age:PClass, data = titanic, family = "binomial")
age_first_class = summary(model_age_pclass)$coefficients["Age:PClass1st", "Pr(>|z|)"]
age_second_class = summary(model_age_pclass)$coefficients["Age:PClass2nd", "Pr(>|z|)"]
age_third_class = summary(model_age_pclass)$coefficients["Age:PClass3rd", "Pr(>|z|)"]
```

Age * PClass interaction Big P values (> 0.05) for Age:PClass1st (0.0724887) and Age:PClass3rd (5.972008×10^{-14}) suggest there is no strong evidence of an interaction between age and those classes passengers.

Low P value ($\ll 0.05$) for Age:PClass2nd (5.8083981×10^{-6}) suggest there strong evidence of an interaction between age and 2nd class of passengers.

```
model_age_sex <- glm(Survived ~ Age:Sex, data = titanic, family = "binomial")
age_female = summary(model_age_sex)$coefficients["Age:Sexfemale", "Pr(>|z|)"]
age_male = summary(model_age_sex)$coefficients["Age:Sexmale", "Pr(>|z|)"]
```

Age * Sex interaction Low P values ($\ll 0.05$) for Age:Sexfemale (7.2489359×10^{-7}) and Age:Sexmale (age_male) suggest there is strong evidence of an interaction between age and the sex of the passenger

Hypothetical of a 55 years old person

We build a linear model to predict Survival with respect to Age, PClass, Sex and their interactions.

```
model <- glm(Survived ~ Age*PClass*Sex, data = titanic, family = "binomial")
male_first <- predict(model, data.frame(Age = 55, PClass = "1st", Sex = "male"), type = "response")
male_second <- predict(model, data.frame(Age = 55, PClass = "2nd", Sex = "male"), type = "response")
male_third <- predict(model, data.frame(Age = 55, PClass = "3rd", Sex = "male"), type = "response")

female_first <- predict(model, data.frame(Age = 55, PClass = "1st", Sex = "female"), type = "response")
female_second <- predict(model, data.frame(Age = 55, PClass = "2nd", Sex = "female"), type = "response")
female_third <- predict(model, data.frame(Age = 55, PClass = "3rd", Sex = "female"), type = "response")
```

- Male
 - 55 years old male of 1st class has a change of 18.5658098% to survive.
 - 55 years old male of 2nd class has a change of 0.2056975% to survive.
 - 55 years old male of 3rd class has a change of 3.5889482% to survive.

- Female
 - 55 years old **female** of **1st** class has a change of 96.0046036% to survive.
 - 55 years old **female** of **2nd** class has a change of 77.6731219% to survive.
 - 55 years old **female** of **3rd** class has a change of 44.5486448% to survive.

Based on previous predictions we argue that a 1st class 55 years old female 96.0046036% was very likely to survive, while a 2nd class 55 years old male (96.0046036%) would have likely died.
