PHUONG Kaggle project DSCI 478

2025-02-19

LOAD DATA SET

```r
# Read Titanic data set
train_df <- read.csv("train.csv",
    stringsAsFactors = FALSE)
test_df <- read.csv("test.csv",
    stringsAsFactors = FALSE)

head(train_df)
```

```
##   PassengerId Survived Pclass
## 1           1        0      3
## 2           2        1      1
## 3           3        1      3
## 4           4        1      1
## 5           5        0      3
## 6           6        0      3
##                                                  Name    Sex Age SibSp Parch
## 1                             Braund, Mr. Owen Harris   male  22     1     0
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1     0
## 3                              Heikkinen, Miss. Laina female  26     0     0
## 4        Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1     0
## 5                            Allen, Mr. William Henry   male  35     0     0
## 6                                    Moran, Mr. James   male  NA     0     0
##             Ticket    Fare Cabin Embarked
## 1        A/5 21171  7.2500              S
## 2         PC 17599 71.2833   C85        C
## 3 STON/O2. 3101282  7.9250              S
## 4           113803 53.1000  C123        S
## 5           373450  8.0500              S
## 6           330877  8.4583              Q
```

```r
head(test_df)
```

```
##   PassengerId Pclass                                         Name    Sex  Age
## 1         892      3                             Kelly, Mr. James   male 34.5
## 2         893      3             Wilkes, Mrs. James (Ellen Needs) female 47.0
## 3         894      2                    Myles, Mr. Thomas Francis   male 62.0
## 4         895      3                             Wirz, Mr. Albert   male 27.0
## 5         896      3 Hirvonen, Mrs. Alexander (Helga E Lindqvist) female 22.0
## 6         897      3                   Svensson, Mr. Johan Cervin   male 14.0
##   SibSp Parch  Ticket    Fare Cabin Embarked
## 1     0     0  330911  7.8292              Q
## 2     1     0  363272  7.0000              S
## 3     0     0  240276  9.6875              Q
## 4     0     0  315154  8.6625              S
## 5     1     1 3101298 12.2875              S
## 6     0     0    7538  9.2250              S
```

CHECK FOR MISSING VALUES

```r
# Count missing values in
# each column
colSums(is.na(train_df))
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0           0           0           0           0         177
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           0           0           0
```

```r
# Remove rows where Age is
# missing
train_df <- train_df[!is.na(train_df$Age),
    ]
```

DOUBLE CHECK TO MAKE SURE THERE ARE NO MORE MISSING VALUES AFTER REMOVING ROWS TO BE DONE CORRECTLY

```r
# Count missing values in
# each column
colSums(is.na(train_df))
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0           0           0           0           0           0
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           0           0           0
```

CONVERT CATEGORICAL VARIABLES TO NUMERIC

```r
# Sex: Male -> 0; Females ->
# 1
train_df$Sex <- ifelse(train_df$Sex ==
    "male", 0, 1)
test_df$Sex <- ifelse(test_df$Sex ==
    "male", 0, 1)

# Embarked: 'C' (Cherbourg)
# -> 0; 'Q' (Queenstown) ->
# 1; 'S' (Southampton) -> 2
train_df$Embarked <- ifelse(train_df$Embarked ==
    "C", 0, ifelse(train_df$Embarked ==
    "Q", 1, 2))
test_df$Embarked <- ifelse(test_df$Embarked ==
    "C", 0, ifelse(test_df$Embarked ==
    "Q", 1, 2))

# Fare: Round to two decimal
# places
train_df$Fare <- round(train_df$Fare,
    2)
test_df$Fare <- round(test_df$Fare,
    2)

# Check the first few rows to
# confirm changes
head(train_df[, c("PassengerId",
    "Survived", "Pclass", "Sex",
    "Age", "SibSp", "Parch", "Fare",
    "Embarked")])
```

```
##   PassengerId Survived Pclass Sex Age SibSp Parch  Fare Embarked
## 1           1        0      3   0  22     1     0  7.25        2
## 2           2        1      1   1  38     1     0 71.28        0
## 3           3        1      3   1  26     0     0  7.92        2
## 4           4        1      1   1  35     1     0 53.10        2
## 5           5        0      3   0  35     0     0  8.05        2
## 7           7        0      1   0  54     0     0 51.86        2
```

SAVE CLEAN DATASET

```r
write.csv(train_df, "clean_train.csv",
    row.names = FALSE)
write.csv(test_df, "clean_test.csv",
    row.names = FALSE)

head(train_df)
```

```
##   PassengerId Survived Pclass
## 1           1        0      3
## 2           2        1      1
## 3           3        1      3
## 4           4        1      1
## 5           5        0      3
## 7           7        0      1
##                                                   Name Sex Age SibSp Parch
## 1                              Braund, Mr. Owen Harris   0  22     1     0
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer)   1  38     1     0
## 3                               Heikkinen, Miss. Laina   1  26     0     0
## 4        Futrelle, Mrs. Jacques Heath (Lily May Peel)   1  35     1     0
## 5                             Allen, Mr. William Henry   0  35     0     0
## 7                              McCarthy, Mr. Timothy J   0  54     0     0
##             Ticket  Fare Cabin Embarked
## 1        A/5 21171  7.25              2
## 2         PC 17599 71.28   C85        0
## 3 STON/O2. 3101282  7.92              2
## 4           113803 53.10  C123        2
## 5           373450  8.05              2
## 7            17463 51.86   E46        2
```

LOADING CLEAN DATASET

```r
# Load the cleaned dataset
train_df <- read.csv("clean_train.csv",
    stringsAsFactors = FALSE)
test_df <- read.csv("clean_test.csv",
    , stringsAsFactors = FALSE)

# Check the first few rows to
# verify correctness
head(train_df)
```

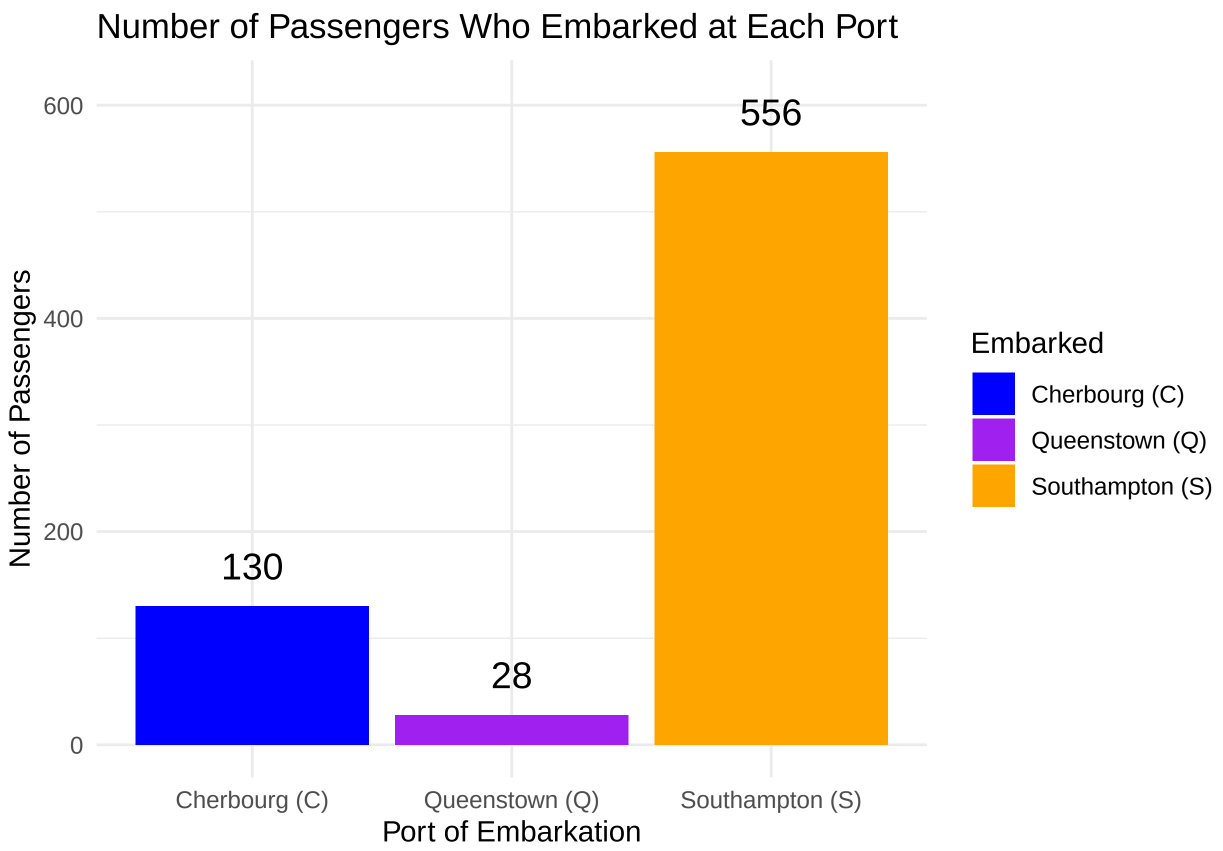
```
##   PassengerId Survived Pclass
## 1           1        0      3
## 2           2        1      1
## 3           3        1      3
## 4           4        1      1
## 5           5        0      3
## 6           7        0      1
##                                                   Name Sex Age SibSp Parch
## 1                              Braund, Mr. Owen Harris   0  22     1     0
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer)   1  38     1     0
## 3                               Heikkinen, Miss. Laina   1  26     0     0
## 4        Futrelle, Mrs. Jacques Heath (Lily May Peel)   1  35     1     0
## 5                             Allen, Mr. William Henry   0  35     0     0
## 6                              McCarthy, Mr. Timothy J   0  54     0     0
##             Ticket  Fare Cabin Embarked
## 1        A/5 21171  7.25              2
## 2         PC 17599 71.28   C85        0
## 3 STON/O2. 3101282  7.92              2
## 4           113803 53.10  C123        2
## 5           373450  8.05              2
## 6            17463 51.86   E46        2
```

```r
colSums(is.na(train_df))
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0           0           0           0           0           0
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           0           0           0
```

DATA EXPLORATORY

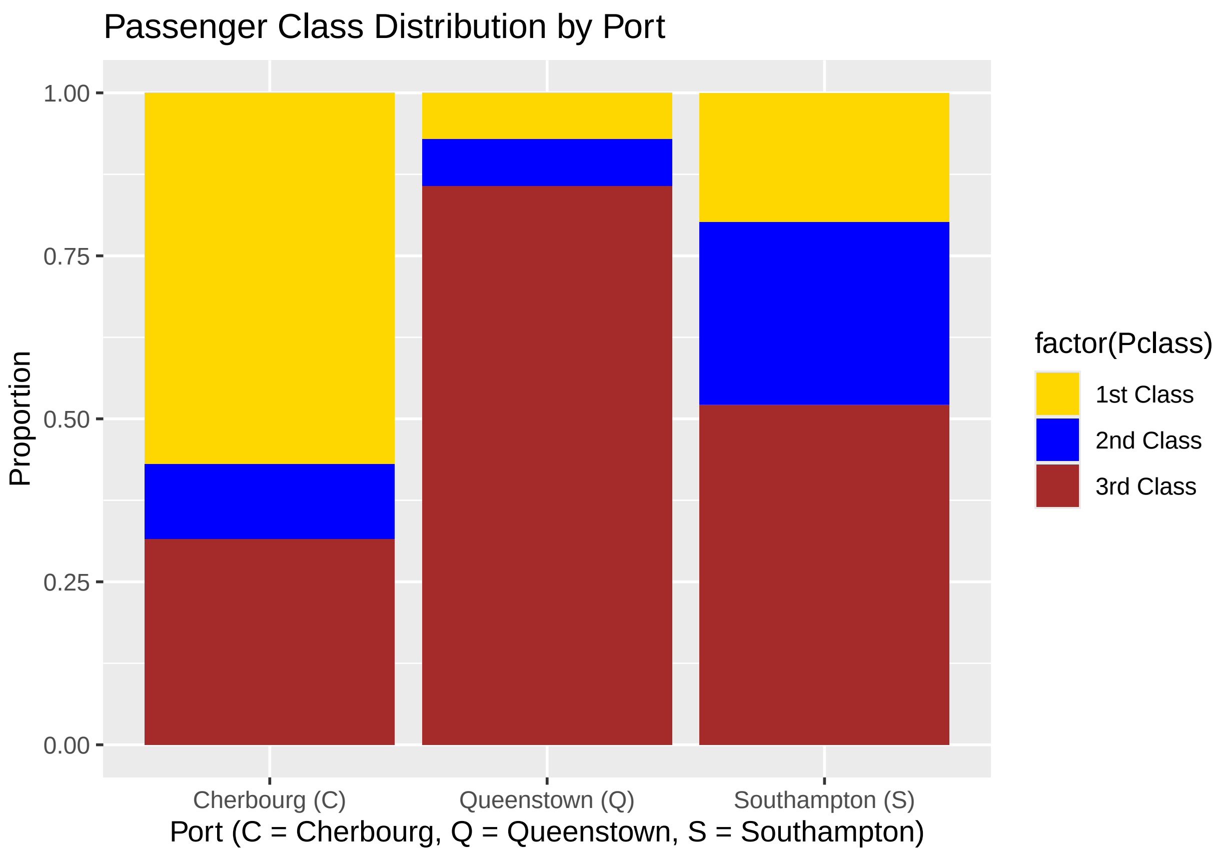
```r
train_df$Embarked <- factor(train_df$Embarked, levels = c(0, 1, 2),
                            labels = c("Cherbourg (C)", "Queenstown (Q)", "Southampton (S)"))


embark_counts <- train_df %>%
  count(Embarked)

ggplot(train_df, aes(x = Embarked, fill = Embarked)) +
  geom_bar() +
  geom_text(stat = "count", aes(label = after_stat(count)), vjust = -1, size = 5) +
  labs(title = "Number of Passengers Who Embarked at Each Port",
       x = "Port of Embarkation",
       y = "Number of Passengers") +
  scale_fill_manual(values = c("blue", "purple", "orange")) +
  ylim(0, max(table(train_df$Embarked)) * 1.1) +  # Extend y-axis for more space
  theme_minimal()
```



This bar chart represents the number of Titanic passengers who embarked at three different ports: Southampton (556 passengers), Cherbourg (130 passengers), and Queenstown (28 passengers). Southampton had the highest number of passengers, as it was the Titanic's primary departure point and a major port in the United Kingdom. Cherbourg, a significant stop in France, saw a moderate number of passengers boarding, while Queenstown, located in Ireland, had the fewest passengers. The clear disparity in embarkation numbers highlights Southampton's importance as a hub for transatlantic travel.

The graph also suggests differences in travel patterns among the three ports. Southampton, as the starting point of the voyage, naturally had the largest group of passengers, while Cherbourg likely served as a key stop for European travelers joining the journey. Queenstown, though an embarkation point, had a significantly smaller number of passengers, indicating that it played a minor role in the overall boarding process. This visualization effectively captures the distribution of embarkation and provides insight into how passenger numbers varied by port.

```
ggplot(train_df, aes(x = Embarked,
    fill = factor(Pclass))) + geom_bar(position = "fill") +
    labs(title = "Passenger Class Distribution by Port",
        x = "Port (C = Cherbourg, Q = Queenstown, S = Southampton)",
        y = "Proportion") + scale_fill_manual(values = c("gold",
    "blue", "brown"), labels = c("1st Class",
    "2nd Class", "3rd Class"))
```



The bar chart “Passenger Class Distribution by Port” illustrates the proportion of First-Class, Second-Class, and Third-Class passengers who embarked from Cherbourg, Queenstown, and Southampton. Cherbourg had the highest share of First-Class passengers, suggesting a wealthier group, while Queenstown was dominated by Third-Class passengers, primarily emigrants seeking opportunities in America. Southampton, the largest departure port, had a more balanced mix but still saw a majority of Third-Class travelers, along with the highest proportion of Second-Class passengers.

This distribution reflects the socioeconomic differences among passengers, which likely influenced survival rates. Cherbourg’s higher proportion of First-Class travelers may explain its relatively better survival outcomes, as wealthier passengers had greater access to lifeboats. Queenstown’s predominantly Third-Class population had lower survival odds due to poorer access and physical barriers on the ship. Meanwhile, Southampton’s diverse mix of classes represents the broader passenger demographic of the Titanic, where Third-Class passengers remained the most vulnerable.

LINEAR REGRESSION MODEL

```
# Convert categorical
# variables to factors
train_df$Pclass <- as.factor(train_df$Pclass)
train_df$Sex <- as.factor(train_df$Sex)
train_df$Embarked <- as.factor(train_df$Embarked)

test_df$Pclass <- as.factor(test_df$Pclass)
test_df$Sex <- as.factor(test_df$Sex)

# Ensure Embarked levels
# match between train and
# test
```

```
test_df$Embarked <- factor(test_df$Embarked,
    levels = levels(train_df$Embarked))

# Convert Survived to a
# factor (classification
# problem)
train_df$Survived <- as.factor(train_df$Survived)

# Define the formula for
# logistic regression
formula <- Survived ~ Pclass +
    Sex + Age + SibSp + Parch +
    Fare + Embarked

# Train a logistic regression
# model
logit_model <- glm(formula, data = train_df,
    family = "binomial")

# Print model summary
print(summary(logit_model))
```

```
##
## Call:
## glm(formula = formula, family = "binomial", data = train_df)
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                1.789595   0.481338   3.718 0.000201 ***
## Pclass2                   -1.199108   0.329081  -3.644 0.000269 ***
## Pclass3                   -2.403087   0.343396  -6.998 2.60e-12 ***
## Sex1                       2.648342   0.222682  11.893  < 2e-16 ***
## Age                       -0.043131   0.008311  -5.190 2.11e-07 ***
## SibSp                     -0.364667   0.129459  -2.817 0.004850 **
## Parch                     -0.062501   0.123969  -0.504 0.614145
## Fare                       0.001484   0.002602   0.570 0.568431
## EmbarkedQueenstown (Q)    -0.821843   0.600874  -1.368 0.171392
## EmbarkedSouthampton (S)   -0.395726   0.274553  -1.441 0.149486
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 964.52  on 713  degrees of freedom
## Residual deviance: 632.91  on 704  degrees of freedom
## AIC: 652.91
##
## Number of Fisher Scoring iterations: 5
```

```
# Make probability
# predictions on training
# data
train_prob <- predict(logit_model,
    train_df, type = "response")
```

```
# Convert probabilities to
# class labels (0 or 1)
train_pred <- as.factor(ifelse(train_prob >
    0.5, 1, 0))

# Evaluate model performance
conf_matrix <- confusionMatrix(train_pred,
    train_df$Survived)

# Print confusion matrix and
# statistics
print(conf_matrix)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 365  83
##          1  59 207
##
##                Accuracy : 0.8011
##                  95% CI : (0.7699, 0.8298)
##     No Information Rate : 0.5938
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.5823
##
##  Mcnemar's Test P-Value : 0.05359
##
##             Sensitivity : 0.8608
##             Specificity : 0.7138
##          Pos Pred Value : 0.8147
##          Neg Pred Value : 0.7782
##              Prevalence : 0.5938
##          Detection Rate : 0.5112
##    Detection Prevalence : 0.6275
##       Balanced Accuracy : 0.7873
##
##        'Positive' Class : 0
##
```

Key Findings from the Logistic Regression Model The logistic regression model identified several factors that played a crucial role in determining whether a passenger survived.

One of the most significant predictors was passenger class (Pclass). The analysis revealed that first-class passengers had the highest survival rates, while third-class passengers faced the highest mortality rates. Specifically, second-class passengers had a 1.20-unit decrease in log-odds of survival compared to first-class passengers (Estimate = -1.199, p = 0.0003), and third-class passengers had a 2.40-unit decrease (Estimate = -2.403, $p < 0.0001$). These findings indicate that social class played a decisive role in survival, likely due to better access to lifeboats for wealthier passengers and the physical layout of the ship, which placed third-class passengers in the lower decks.

Another highly significant factor was sex. The model found that being female greatly increased the probability of survival (Estimate = 2.648, $p < 0.0001$). This supports historical accounts that women were prioritized in the evacuation process, following the “women and children first” protocol.

Age also influenced survival, though to a lesser extent. The model showed that each additional year of age slightly decreased survival odds (Estimate = -0.043, p = 0.0002). This suggests that younger passengers had a better chance of survival, possibly because they were given priority or were more physically capable of reaching lifeboats.

The number of family members traveling together also had an effect. Having more siblings or spouses aboard was associated with lower survival odds (Estimate = -0.365, p = 0.0049). This could be because families tried to stay together, which may have delayed evacuation efforts. However, the number of parents or children aboard did not significantly impact survival (Estimate = -0.062, p = 0.614). One possible explanation is that families with young children were prioritized, but beyond that, the presence of parents or children did not make a meaningful difference.
Interestingly, ticket fare was not a significant predictor of survival once passenger class was accounted for (Estimate = 0.0015, p = 0.568). While one might assume that higher ticket prices would indicate a higher survival probability, the model suggests that it was class itself, rather than fare, that mattered most.

Similarly, embarkation point (Cherbourg, Queenstown, or Southampton) did not have a significant effect on survival. Although passengers from Cherbourg showed slightly better survival rates, the difference was not statistically significant, suggesting that where a passenger boarded was not a crucial factor in survival.

Model Performance Evaluation To assess the accuracy of the logistic regression model, a confusion matrix was generated. The model correctly classified 80.11% of passengers, demonstrating strong predictive performance. The p-value (< 2e-16) confirmed that the model performed significantly better than random guessing.

A more detailed breakdown of the results showed that the model was better at predicting non-survivors than survivors. The sensitivity (recall for non-survivors) was 86.08%, meaning the model correctly identified 86.08% of those who did not survive. However, the specificity (recall for survivors) was lower, at 71.38%, indicating that the model was slightly less effective at correctly identifying survivors. The balanced accuracy, which averages sensitivity and specificity, was 78.73%, suggesting a well-performing model overall.

Conclusion This analysis confirms that passenger class, sex, and age were the most important determinants of survival on the Titanic. Women and first-class passengers had the highest survival rates, while older individuals and those in third-class were less likely to survive. Having more family members aboard was generally disadvantageous, particularly for those traveling with siblings or spouses. Other factors, such as fare price and embarkation location, had little to no impact on survival.

Ultimately, this study highlights the harsh reality of social class disparities and the historical prioritization of women and children during the Titanic disaster. By leveraging data analysis techniques, we can gain a deeper understanding of the factors that influenced survival and improve predictive modeling for future historical and real-world applications.