# HW 1

## Use Quarto to Summarize the Titanic Dataset

何淑雯 RE6137012

2025-03-07

## Table of contents



## 1.Data Introduction

```
# Load Dataset
titanic<-read.csv("C:/Users/ACER/Downloads/titanic.csv")
View(titanic)
```

The dataset is about the 891 passengers of the Titanic. It contains 12 Variables which include Passenger ID (PassengerId), Survival Status (Survived), Passenger Class (Pclass), Name, Sex, Age, Number of Siblings/Spouse Aboard (SibSp), Number of Parents/Children Aboard (Parch), Ticket Number (Ticket), Ticket Fare (Fare), Cabin Number (Cabin), and Port of Embarkation (Embarked).

```
# Data Structure
library(knitr)
titanic.class <- data.frame(Variable = names(titanic), Class = sapply(titanic, class))
kable(titanic.class, caption = "Class of Each Variable in Titanic Dataset")
```

Table 1: Class of Each Variable in Titanic Dataset

| | Variable | Class |
|---|---|---|
| PassengerId | PassengerId | integer |
| Survived | Survived | integer |
| Pclass | Pclass | integer |
| Name | Name | character |
| Sex | Sex | character |
| Age | Age | numeric |
| SibSp | SibSp | integer |
| Parch | Parch | integer |
| Ticket | Ticket | character |
| Fare | Fare | numeric |
| Cabin | Cabin | character |
| Embarked | Embarked | character |

From Table 1, it was seen that PassengerId, Survived, Pclass, SibSp, Parch are all integers. Name, Sex, Ticket, Cabin, Embarked are characters. Meanwhile, Age and Fare are numeric. There are 2 levels (binary) for Survived: 0 indicates the passanger didn't survive and 1 indicates that the passenger survived. There are 3 levels for Pclass: 1 to 3 depending on which socio-economic status the passenger belongs to.

## 2.Data Organization

```
# Remove Irrelevant Variables
titanic.new <- titanic[, !(names(titanic) %in% c("PassengerId", "Name", "Ticket", "Cabin"))]
View(titanic.new)
```

Variables such as PassengerId, Name, Ticket, and Cabin are exempted from the analysis because the variables are unique characters for each observation. Therefore, it will be irrelevant to be analyzed. Also, observations with at least one missing value are omitted, so that it will not affect the analysis process. Total of 714 data are proceeded for analysis process.

```
# Variables Classification
titanic_variables <- data.frame(
  Variable = c("Survived", "Pclass", "Sex", "Age",
               "SibSp", "Parch", "Fare", "Embarked"),
  Type = c("Catagorical", "Categorical", "Categorical", "Numeric", "Numeric", "Numeric", "Numeric", "Cat
  Category = c("Nominal", "Ordinal", "Nominal", "Continuous","Discrete", "Discrete","Continuous", "Nomir
)
kable(titanic_variables, caption = "Titanic Dataset Variable Classification")
```

Table 2: Titanic Dataset Variable Classification

| Variable | Type | Category |
|---|---|---|
| Survived | Catagorical | Nominal |
| Pclass | Categorical | Ordinal |
| Sex | Categorical | Nominal |
| Age | Numeric | Continuous |
| SibSp | Numeric | Discrete |
| Parch | Numeric | Discrete |
| Fare | Numeric | Continuous |
| Embarked | Categorical | Nominal |

From Table 2, it can be seen that the variables can be grouped into 3 categories: Ordinal, Nominal, Cardinal (Discrete/Continuous). Survived, Sex and Embarked are Nominal because those are qualitative variables that represent categories without any inherent order. Pclass is Ordinal because it is a categorical variable with a meaningful order. Lastly, Age, SibSp, Parch, Fare are Cardinal because the variables are quantitative and it can be either measured (continuous) or counted (dicrete).

## 3.Data Summary

```
# Summary Statistics for Nominal and Ordinal Variables
summary_table1 <- data.frame(
  Category = c("Survived (0)", "Survived (1)",
               "Pclass (1)", "Pclass (2)", "Pclass (3)",
               "Sex (Female)", "Sex (Male)",
               "Embarked (Unknown)", "Embarked (C)", "Embarked (Q)", "Embarked (S)"),
```

```
  Count = c(table(titanic.new$Survived)[1], table(titanic.new$Survived)[2],
            table(titanic.new$Pclass)[1], table(titanic.new$Pclass)[2], table(titanic.new$Pclass)[3],
            table(titanic.new$Sex)[1], table(titanic.new$Sex)[2],
            table(titanic.new$Embarked)[1], table(titanic.new$Embarked)[2],
            table(titanic.new$Embarked)[3], table(titanic.new$Embarked)[4])
)

kable(summary_table1, caption = "Titanic Dataset Summary for Nominal and Ordinal Variables")
```

Table 3: Titanic Dataset Summary for Nominal and Ordinal Variables

| Category | Count |
|---|---|
| Survived (0) | 549 |
| Survived (1) | 342 |
| Pclass (1) | 216 |
| Pclass (2) | 184 |
| Pclass (3) | 491 |
| Sex (Female) | 314 |
| Sex (Male) | 577 |
| Embarked (Unknown) | 2 |
| Embarked (C) | 168 |
| Embarked (Q) | 77 |
| Embarked (S) | 644 |

From Table 3, it can be stated that there are 549 passengers who didn' t survive. 491 passengers belong to the 3rd class, 216 passengers belong to the first class and 184 passengers belong to the second class. 577 of the passengers on board are Male. 644 passengers embarked from Southampton (S), 168 embarked from Cherbourg (C) and 77 passengers embarked from Queenstown (Q), meanwhile the embarkation point of the other 2 passengers are unknown.

```
# Summary Statistics for Cardinal Variables
library(fBasics)
sum<- round(cbind(basicStats(titanic.new$Age), basicStats(titanic.new$SibSp),
basicStats(titanic.new$Parch), basicStats(titanic.new$Fare)),3)
colnames(sum) <- c("Age","SibSp","Parch","Fare")
#range
range <- sum[4,]-sum[3,] #max-min
rownames(range) <- "Range"
#IQR
IQR <- sum[6,]-sum[5,] #Q3-Q1
rownames(IQR) <- "IQR"
summary_table2<- rbind(sum[3,],sum[4,],sum[7,],sum[8,],sum[13,],sum[14,],
range,sum[5,],sum[6,],IQR,sum[15,])
summary_table2
```

```
            Age SibSp Parch     Fare
```

Minimum 0.420 0.000 0.000 0.000 Maximum 80.000 8.000 6.000 512.329 Mean 29.699 0.523 0.382 32.204 Median 28.000 0.000 0.000 14.454 Variance 211.019 1.216 0.650 2469.437 Stdev 14.526 1.103 0.806 49.693 Range 79.580 8.000 6.000 512.329 1. Quartile 20.125 0.000 0.000 7.910 3. Quartile 38.000 1.000 0.000 31.000 IQR 17.875 1.000 0.000 23.090 Skewness 0.387 3.683 2.740 4.771

```
kable(summary_table2, caption = "Titanic Dataset Summary for Cardinal Variables")
```

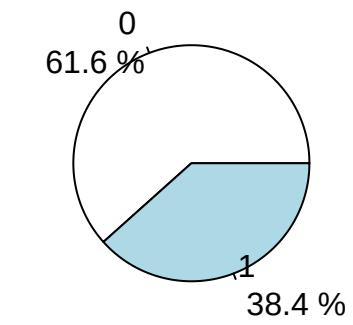
Table 4: Titanic Dataset Summary for Cardinal Variables

| | Age | SibSp | Parch | Fare |
|---|---|---|---|---|
| Minimum | 0.420 | 0.000 | 0.000 | 0.000 |
| Maximum | 80.000 | 8.000 | 6.000 | 512.329 |
| Mean | 29.699 | 0.523 | 0.382 | 32.204 |
| Median | 28.000 | 0.000 | 0.000 | 14.454 |
| Variance | 211.019 | 1.216 | 0.650 | 2469.437 |
| Stdev | 14.526 | 1.103 | 0.806 | 49.693 |
| Range | 79.580 | 8.000 | 6.000 | 512.329 |
| 1. Quartile | 20.125 | 0.000 | 0.000 | 7.910 |
| 3. Quartile | 38.000 | 1.000 | 0.000 | 31.000 |
| IQR | 17.875 | 1.000 | 0.000 | 23.090 |
| Skewness | 0.387 | 3.683 | 2.740 | 4.771 |

From the table above, the distribution of the passengers' age centered on around 28-30 years old, with mean of 29-30 years old and median of 28 years old. The range of the eldest to the youngest is around 79-80 years old. The eldest passeger is 80 years old and the youngest passenger is under 1 year old. Most of the passengers are alone, they didn't bring any siblings/spouse or children/parents. This can be seen from the median of both SibSp and Parch which are 0. As for the ticket fare, the passengers paid 34.70 dollars on average, with the most expensive cost of 512.329 dollars.

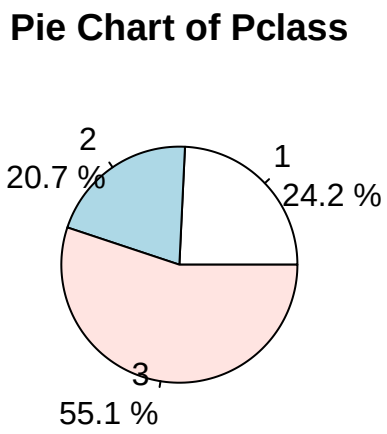
## 4.Data Visualization

```
# Pie Chart for Nominal and Ordinal Variables
survived_counts <- table(titanic.new$Survived)
survived_percent <- round(100 * survived_counts / sum(survived_counts), 1)
labels_survived <- paste(names(survived_counts), "\n", survived_percent, "%")
pie(survived_counts, labels = labels_survived, main="Pie Chart of Survived")
```



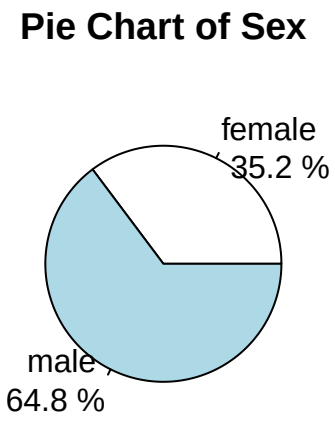
From the Pie chart above, more than half of the passengers didn’ t survive in the Titanic tragedy (shown in white) and 38.4% of the passengers survive (shown in blue).

```
pclass_counts <- table(titanic.new$Pclass)
pclass_percent <- round(100 * pclass_counts / sum(pclass_counts), 1)
labels_pclass <- paste(names(pclass_counts), "\n", pclass_percent, "%")
pie(pclass_counts, labels = labels_pclass, main="Pie Chart of Pclass")
```



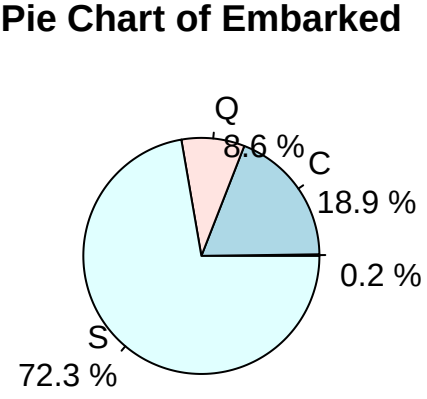

The pie chart above shows that 55.1% of the passengers belongs to the 3rd class, followed by 1st class and 2nd class.

```
sex_counts <- table(titanic.new$Sex)
sex_percent <- round(100 * sex_counts / sum(sex_counts), 1)
labels_sex <- paste(names(sex_counts), "\n", sex_percent, "%")
pie(sex_counts, labels = labels_sex, main="Pie Chart of Sex")
```



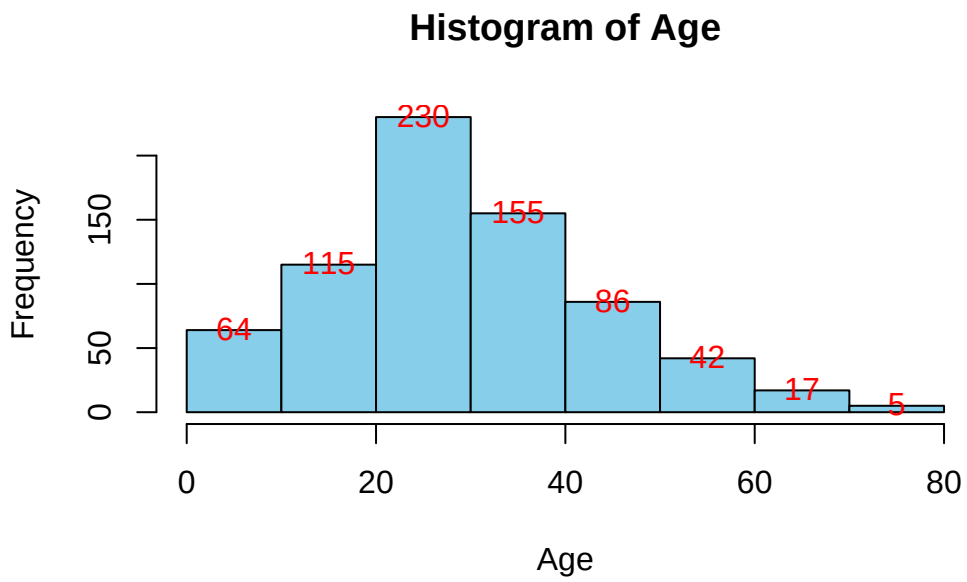

The Pie chart above shows that most of the passengers are male. In fact, the percentage of Male is almost 2 times the percentage of Female on board.

```
embarked_counts <- table(titanic.new$Embarked)
embarked_percent <- round(100 * embarked_counts / sum(embarked_counts), 1)
labels_embarked <- paste(names(embarked_counts), "\n", embarked_percent, "%")
pie(embarked_counts, labels = labels_embarked, main="Pie Chart of Embarked")
```



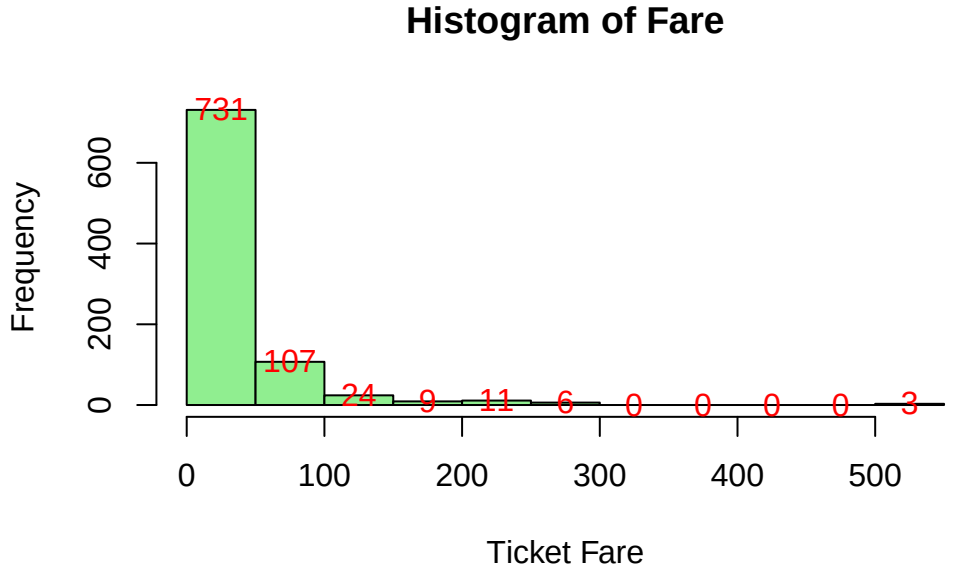

From the pie chart above, it can be concluded that almost three-fourth of the passengers embarked from Southampton, followed by Cherbourg and Queenstown. The 0.2% represents missing data values where the embarkation port of the passenger is unknown.

```
# Histogram and Bar Chart for Cardinal Variables
hist_data <- hist(titanic.new$Age, main = "Histogram of Age", xlab = "Age", col="skyblue")
text(hist_data$mids, hist_data$counts + 1, labels=hist_data$counts, col="red", cex=1.0)
```



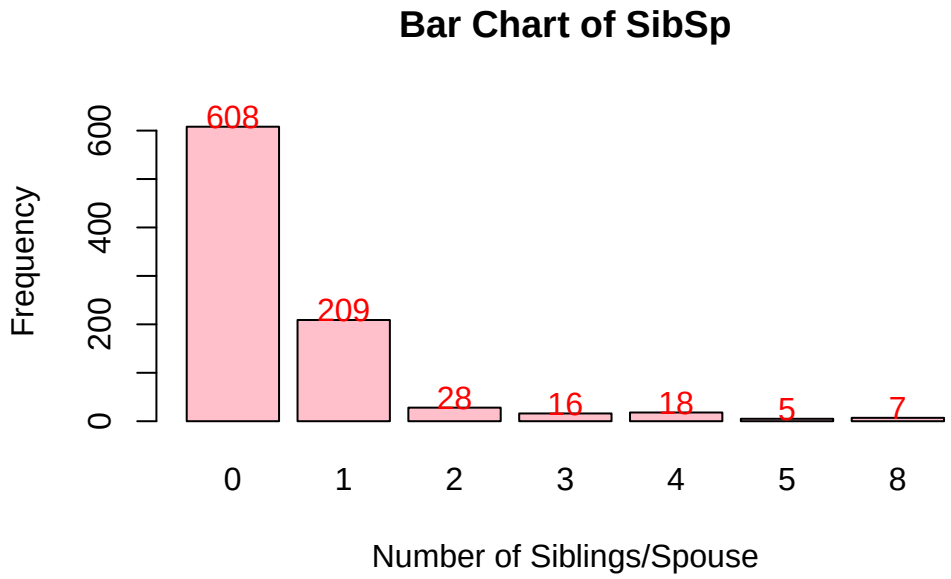

From the histogram shown above, it is obvious that most of the passengers are aged 20-30 years old with a total of 230 people, followed by 30-40 years old and 10-20 years old. The least populated age group is 70-80 years old with only 5 people.

```
hist_fare <- hist(titanic.new$Fare, main = "Histogram of Fare", xlab = "Ticket Fare", col="lightgreen")
text(hist_fare$mids, hist_fare$counts + 1, labels=hist_fare$counts, col="red", cex=1.0)
```



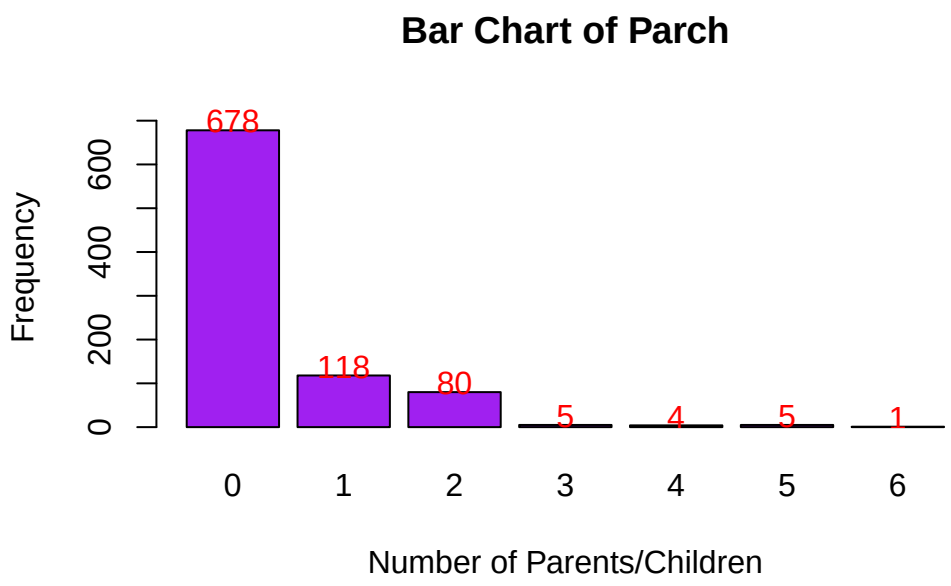
From the Histogram of Fare, it can be seen that most passengers paid 0-50 dollars for their ticket. Only 3 people paid more than 500 dollars for their tickets, while the rest paid less than 300 dollars. No tickets were sold between 301-499 dollars.

```
sibsp_counts <- table(titanic.new$SibSp)
barplot_heights <- barplot(sibsp_counts, main="Bar Chart of SibSp", xlab="Number of Siblings/Spouse", yl
text(barplot_heights, sibsp_counts + 20, labels = sibsp_counts, col="red", cex=1.0)
```




According to the Bar Chart of SibSp, most people didn't bring any siblings/spouse aboard, but if they did, they're most likely to bring only 1 sibling/spouse.

```
parch_counts <- table(titanic.new$Parch)
barplot_heights <- barplot(parch_counts, main="Bar Chart of Parch", xlab="Number of Parents/Children", y
text(barplot_heights, parch_counts + 20, labels = parch_counts, col="red", cex=1.0)
```




According to the bar chart above, like Siblings/Spouse, most people didn’ t bring Parents/Children aboard. In fact more than three-fourth of the population didn’ t bring parents/children with them.