Milestone_5

Charles James

2023-10-15

```
titanic = read.csv("titanic.csv")
colnames(titanic)
```

```
##  [1] "PassengerId" "Survived"    "Pclass"      "Name"        "Sex"
##  [6] "Age"         "SibSp"       "Parch"       "Ticket"      "Fare"
## [11] "Cabin"       "Embarked"
```

```
dim(titanic)
```

```
## [1] 891  12
```

```
age = titanic$Age
```

Age Analysis

```
age_mean <- mean(age, na.rm=TRUE)
age_mean
```

```
## [1] 29.69912
```

```
age_median <- median(age, na.rm=TRUE)
age_median
```

```
## [1] 28
```

```
age_variance <- var(age, na.rm=TRUE)
age_variance
```

```
## [1] 211.0191
```

```
age_sd <- sd(age, na.rm=TRUE)
age_sd
```

```
## [1] 14.5265
```

The Age histogram mostly resembles the normal distribution with minor skewing to the right. These large outlier ages will bring the mean slightly higher than the median. The mean age is 29.7 while the median age is 28 becuase of the minor outliers. The standard deviation of 14 means that the majority of people on board were between the age of 16 - 44 since the mean is approximately 30. This shows that the age values are spread out from one another.

Fare Analysis

```
fare = titanic$Fare
fare_mean = mean(fare)
fare_mean
```

```
## [1] 32.20421
```

```
fare_median = median(fare)
fare_median
```

```
## [1] 14.4542
```

```
fare_variance = var(fare)
fare_variance
```

```
## [1] 2469.437
```

```
fare_sd = sd(fare)
fare_sd
```

```
## [1] 49.69343
```

The fare data is also skewed right like the age histogram. Here though the outlier fares are significantly larger than the rest of the data. This is why the mean is more than two times the median. The standard deviation of 49 when the mean is 32 is really high and suggests that the data is heavily spread out. Alot of this is cause by the one trip over $500.