EXPLORATORY DATA ANALYSIS

KERAMA EZEKIEL JAMES

2023-05-20

##STEP 1##  
#Load Library#  
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.2.3

library(dplyr)

## Warning: package 'dplyr' was built under R version 4.2.3

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(readxl)

## Warning: package 'readxl' was built under R version 4.2.3

df<- read\_excel("C:/Users/Alif computer/Desktop/tested titanic.xlsx")  
  
# preview data#  
  
head(df)

## # A tibble: 6 × 12  
## PassengerId Survived Pclass Name Sex Age SibSp Parch Ticket Fare Cabin  
## <dbl> <dbl> <dbl> <chr> <chr> <dbl> <dbl> <dbl> <chr> <dbl> <chr>  
## 1 892 0 3 Kelly,… male 34.5 0 0 330911 7.83 <NA>   
## 2 893 1 3 Wilkes… fema… 47 1 0 363272 7 <NA>   
## 3 894 0 2 Myles,… male 62 0 0 240276 9.69 <NA>   
## 4 895 0 3 Wirz, … male 27 0 0 315154 8.66 <NA>   
## 5 896 1 3 Hirvon… fema… 22 1 1 31012… 12.3 <NA>   
## 6 897 0 3 Svenss… male 14 0 0 7538 9.22 <NA>   
## # … with 1 more variable: Embarked <chr>

summary(df)

## PassengerId Survived Pclass Name   
## Min. : 892.0 Min. :0.0000 Min. :1.000 Length:418   
## 1st Qu.: 996.2 1st Qu.:0.0000 1st Qu.:1.000 Class :character   
## Median :1100.5 Median :0.0000 Median :3.000 Mode :character   
## Mean :1100.5 Mean :0.3636 Mean :2.266   
## 3rd Qu.:1204.8 3rd Qu.:1.0000 3rd Qu.:3.000   
## Max. :1309.0 Max. :1.0000 Max. :3.000   
##   
## Sex Age SibSp Parch   
## Length:418 Min. : 0.17 Min. :0.0000 Min. :0.0000   
## Class :character 1st Qu.:21.00 1st Qu.:0.0000 1st Qu.:0.0000   
## Mode :character Median :27.00 Median :0.0000 Median :0.0000   
## Mean :30.27 Mean :0.4474 Mean :0.3923   
## 3rd Qu.:39.00 3rd Qu.:1.0000 3rd Qu.:0.0000   
## Max. :76.00 Max. :8.0000 Max. :9.0000   
## NA's :86   
## Ticket Fare Cabin Embarked   
## Length:418 Min. : 0.000 Length:418 Length:418   
## Class :character 1st Qu.: 7.896 Class :character Class :character   
## Mode :character Median : 14.454 Mode :character Mode :character   
## Mean : 35.627   
## 3rd Qu.: 31.500   
## Max. :512.329   
## NA's :1

str(df)

## tibble [418 × 12] (S3: tbl\_df/tbl/data.frame)  
## $ PassengerId: num [1:418] 892 893 894 895 896 897 898 899 900 901 ...  
## $ Survived : num [1:418] 0 1 0 0 1 0 1 0 1 0 ...  
## $ Pclass : num [1:418] 3 3 2 3 3 3 3 2 3 3 ...  
## $ Name : chr [1:418] "Kelly, Mr. James" "Wilkes, Mrs. James (Ellen Needs)" "Myles, Mr. Thomas Francis" "Wirz, Mr. Albert" ...  
## $ Sex : chr [1:418] "male" "female" "male" "male" ...  
## $ Age : num [1:418] 34.5 47 62 27 22 14 30 26 18 21 ...  
## $ SibSp : num [1:418] 0 1 0 0 1 0 0 1 0 2 ...  
## $ Parch : num [1:418] 0 0 0 0 1 0 0 1 0 0 ...  
## $ Ticket : chr [1:418] "330911" "363272" "240276" "315154" ...  
## $ Fare : num [1:418] 7.83 7 9.69 8.66 12.29 ...  
## $ Cabin : chr [1:418] NA NA NA NA ...  
## $ Embarked : chr [1:418] "Q" "S" "Q" "S" ...

## STEP 2: DATA PREPARATION##  
#Survived#  
df$Survived <- ifelse(df$Survived==1,"Yes","No")  
df$Survived <- as.factor(df$Survived)  
head(df)

## # A tibble: 6 × 12  
## PassengerId Survived Pclass Name Sex Age SibSp Parch Ticket Fare Cabin  
## <dbl> <fct> <dbl> <chr> <chr> <dbl> <dbl> <dbl> <chr> <dbl> <chr>  
## 1 892 No 3 Kelly,… male 34.5 0 0 330911 7.83 <NA>   
## 2 893 Yes 3 Wilkes… fema… 47 1 0 363272 7 <NA>   
## 3 894 No 2 Myles,… male 62 0 0 240276 9.69 <NA>   
## 4 895 No 3 Wirz, … male 27 0 0 315154 8.66 <NA>   
## 5 896 Yes 3 Hirvon… fema… 22 1 1 31012… 12.3 <NA>   
## 6 897 No 3 Svenss… male 14 0 0 7538 9.22 <NA>   
## # … with 1 more variable: Embarked <chr>

##Embarked  
df$Embarked <- ifelse(df$Embarked=="S","Southampton", ifelse(df$Embarked=="C","Cherbourg", "Queenstown"))  
df$Embarked <- as.factor(df$Embarked)  
head(df)

## # A tibble: 6 × 12  
## PassengerId Survived Pclass Name Sex Age SibSp Parch Ticket Fare Cabin  
## <dbl> <fct> <dbl> <chr> <chr> <dbl> <dbl> <dbl> <chr> <dbl> <chr>  
## 1 892 No 3 Kelly,… male 34.5 0 0 330911 7.83 <NA>   
## 2 893 Yes 3 Wilkes… fema… 47 1 0 363272 7 <NA>   
## 3 894 No 2 Myles,… male 62 0 0 240276 9.69 <NA>   
## 4 895 No 3 Wirz, … male 27 0 0 315154 8.66 <NA>   
## 5 896 Yes 3 Hirvon… fema… 22 1 1 31012… 12.3 <NA>   
## 6 897 No 3 Svenss… male 14 0 0 7538 9.22 <NA>   
## # … with 1 more variable: Embarked <fct>

#Converting categorical attributes from int to factor  
  
df$Pclass <- as.factor(df$Pclass)  
df$SibSp <- as.factor(df$SibSp)  
df$Parch <- as.factor(df$Parch)  
head(df)

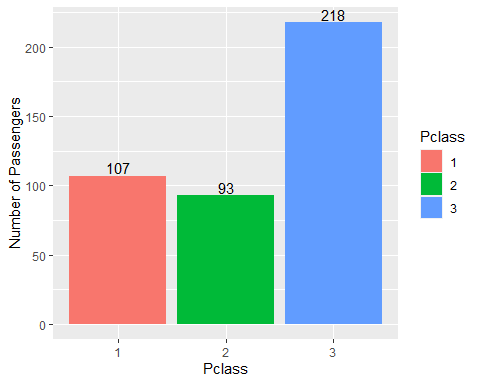
## # A tibble: 6 × 12  
## PassengerId Survived Pclass Name Sex Age SibSp Parch Ticket Fare Cabin  
## <dbl> <fct> <fct> <chr> <chr> <dbl> <fct> <fct> <chr> <dbl> <chr>  
## 1 892 No 3 Kelly,… male 34.5 0 0 330911 7.83 <NA>   
## 2 893 Yes 3 Wilkes… fema… 47 1 0 363272 7 <NA>   
## 3 894 No 2 Myles,… male 62 0 0 240276 9.69 <NA>   
## 4 895 No 3 Wirz, … male 27 0 0 315154 8.66 <NA>   
## 5 896 Yes 3 Hirvon… fema… 22 1 1 31012… 12.3 <NA>   
## 6 897 No 3 Svenss… male 14 0 0 7538 9.22 <NA>   
## # … with 1 more variable: Embarked <fct>

df$Age[is.na(df$Age)] <- round(median(df$Age, na.rm = TRUE))  
  
colSums(is.na(df))

## PassengerId Survived Pclass Name Sex Age   
## 0 0 0 0 0 0   
## SibSp Parch Ticket Fare Cabin Embarked   
## 0 0 0 1 327 0

## STEP THREE: DATA ANALYSIS  
  
#Univariate analysis.  
#Pclass  
ggplot(data=df, aes(x=Pclass, fill = Pclass)) +   
 geom\_bar(position = "dodge") +   
 geom\_text(stat='count', aes(label=..count..), position = position\_dodge(0.9),vjust=-0.2) +  
 ylab("Number of Passengers")

## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.  
## ℹ Please use `after\_stat(count)` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.



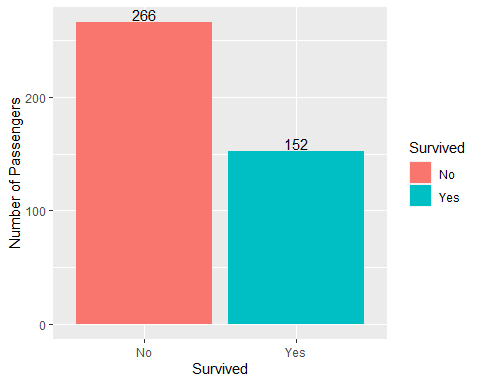
Explanation of the bar plot of the number of passengers against pclass.

~From the bar plot,pclass 3 recorded the highest number of passengers of 218 as compared to pclasses 1 and 2.

~From the bar plot pclass 1 recorded the second highest number of passengers of 107.

~From the bar plot of the number of passengers against pclass,pclass 2 recoded the least number of passengers.It recoded only 93 passengers.

#Survived :  
ggplot(data=df, aes(x=Survived, fill = Survived)) +   
geom\_bar(position = "dodge") +   
 geom\_text(stat='count', aes(label=..count..), position = position\_dodge(0.9),vjust=-0.2) +  
 ylab("Number of Passengers")



Explanation of the bar plot of the survived population.

~From studying the bar plot it was ascertained that from the entire population in the Titanic,a large portion of it perished as just a small number survived.

~From the bar plot those who never survived recorded a population of 266 whereas those who survived recorded a population of 152 passengers.

~From the findings we can therefore say that despite a large population perished there were still those who managed to survive the accident.

#Age  
ggplot(data=df, aes(x=Age,)) +   
 geom\_histogram(binwidth = 5) +  
 xlab("Age")



Explanation of the histogram of age variable.

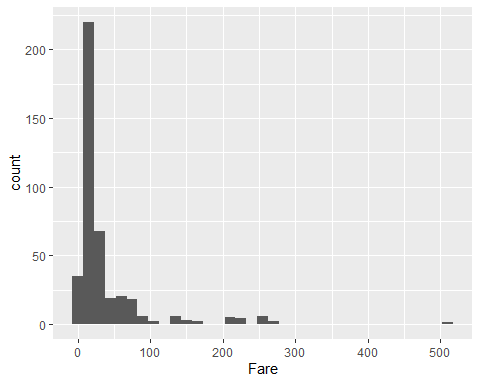
~From the histogram it is seen that as the age was increasing there was a none uniform corresponding increase in the population until the attainment of the climax or limit of 150.

~After the reaching the highest limit of 150 the population started to decrease uniformly with an increase in age.

~From the histogram we can say that as the age increases as well until the highest limit is attained after which even if the age will be increasing the population will be dropping continually.

#Fare  
ggplot(data=df, aes(x=Fare,)) +   
 geom\_histogram(binwidth = 15) +  
 xlab("Fare")

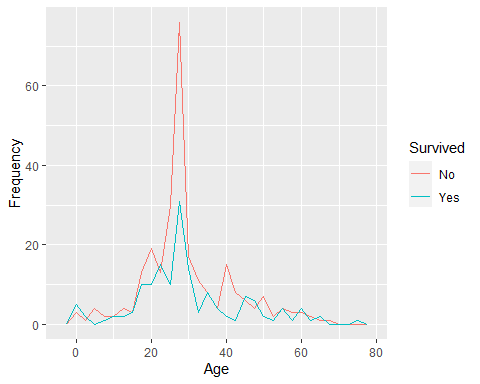
## Warning: Removed 1 rows containing non-finite values (`stat\_bin()`).



Explanation of the histogram of the count against the fare.

~From the histogram plot it is seen that as the fare was increasing the count wasas well increasing until the attainment of the highest limit of about 217.

# Bivariate Analysis  
#survived and Age  
ggplot(df) + geom\_freqpoly(mapping = aes(x = Age, color = Survived), binwidth = 2.5) +  
 ylab("Frequency")



Explanation of the freqpoly frequency against age.

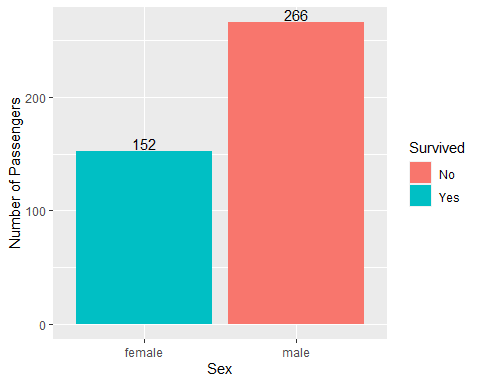
~From the freqpoly plot the frequency was increasing with the increase in age.

~From the freqpoly plot the rate of the population of those who survived was increasing uniformly until it attained a maximum of about 32 after which there was a continuous decrease in the population.

~From the freqpoly plot the rate of the population of those who never survived was increasing uniformly until it attained a maximum of about 76.

~The trend of the population which did not survive was higher than the population of those who survived throughout the period

#Survived and Sex  
ggplot(df, aes(x=Sex,fill=Survived))+ geom\_bar(position = "dodge") + geom\_text(stat='count',aes(label=..count..),position = position\_dodge(0.9),vjust=-0.2) +  
 ylab("Number of Passengers") + xlab("Sex")



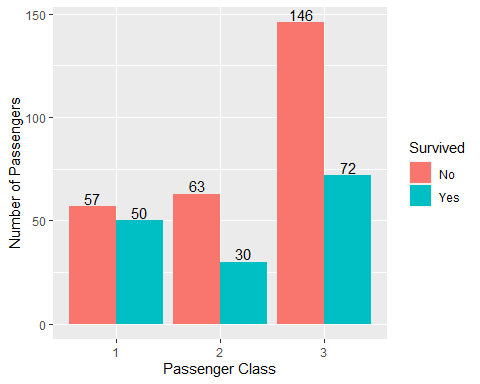
Explanation of the bar plot of passengers against sex.

~From the bar plot it is seen that the number of passengers whose survived the Titanic accident were 152 females.

~From the bar plot it is either seen that the number of males who perished in the accident was 266 greater than that of those females who survived the accident.

~From the bar plot it is seen that the probabity of a male surviving the Titanic accident was lesser than the probability of a female surviving the accident.

#Survived and Pclass  
ggplot(df, aes(x=Pclass,fill=Survived))+ geom\_bar(position = "dodge") + geom\_text(stat='count',aes(label=..count..),position = position\_dodge(0.9),vjust=-0.2) +  
 ylab("Number of Passengers") + xlab("Passenger Class")



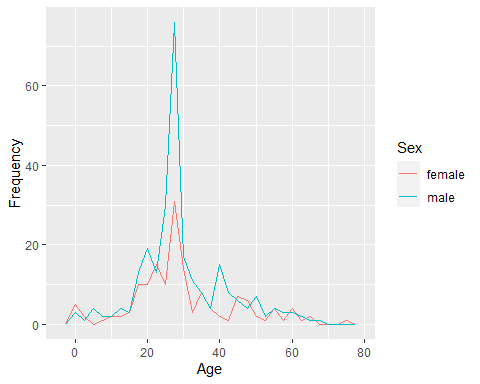
Explanation of the bar plot of the number of passengers against the passenger class.

~From the bar plot it is seen that in passenger class 1 there was a small difference between the survived population to that of the population which did not survive.

~From the bar plot it is seen that in the passenger class 2 the population of those who never survived was twice the population of those who survived the accident.

~From the bar plot it is seen that in the passenger class 3 the population of those who never survived the accident was again twice the population of those who survived the accident.

#Age and sex  
ggplot(df) + geom\_freqpoly(mapping = aes(x = Age, color = Sex), binwidth = 2.5) +  
 ylab("Frequency")



Explanation of the freqpoly of frequency against age.

~From the freqpoly it is seen that as the age increases the frequency increases as well sequencially

#Age and Pclass  
ggplot(df) + geom\_freqpoly(mapping = aes(x = Age, color = Pclass), binwidth = 2.5) +  
 ylab("Frequency")

