###I am choosing R studio to find the insights for a given data, will explain at every step what I am doing######

###setting working directory

setwd("C:/Users/ddeep/OneDrive/Desktop/Business Analyst course/Open Projects/Ozibook Project")

### calling usefull libraries

library(readr)

library(dplyr)

library(plyr)

library(ggplot2)

library(Metrics)

library(statip)

library(tidyverse)

library(tidyr)

###Importing the file to R studio

data = read\_csv("vgsales.csv")

####lets understand the data

head(data)

view(data)

str(data)

summary(data)

dim(data)

table(data$Genre)

table(data$Platform)

###Cleaning data

##changing the data type

data$Year= as.numeric(data$Year)

data$Platform = as.factor(data$Platform)

data$Genre = as.factor(data$Genre)

###lets check is there any NA values

colSums(is.na(data))

###we have NA values in column 'year' of 271 which is less than 5% of total entries

##handling with NA values

data= na.omit(data)

###data looks good to go

### checking the market share in percentage of genre

Genre = data |> group\_by(Genre) |> summarise(count = n()) |> mutate(Percentage = round(count/sum(count)\*100,2)) |> arrange(desc(count))

Genre

####lets check the top three Genre

head(Genre, 4)

#####@@@@@'Action' with 19.9%, 'Sports' with 14.1% 'Role playing' with 9.01% are top three Genre@@@@#######

#lets look at the market distribuytion of genre on graph

ggplot((Genre), aes(x=Genre, y= count))+ geom\_col (fill = 'blue')+ labs(title= "Genre Dsitribution", x= "Genre", y= "Count")

##Lets look at the total gobal sales w.r.t Genre

Genre\_sale = data |> group\_by(Genre) |> summarise(Totalsales = sum(Global\_Sales))|> mutate(percentage\_sale = Totalsales/sum(Totalsales)\*100) |> arrange(desc(percentage\_sale))

Genre\_sale

####Top 3 genre with maximum global sales and their percenatge are -

head(Genre\_sale, 3)

###Lets look this pragphically

ggplot(Genre\_sale, aes(x= Genre, y = Totalsales))+ geom\_col (fill = 'Orange')+ labs(title= 'Global Sales by Genre', x='Genre', y = 'Total Global sales')

######lets now check the sales w.r.t game released Years

sales\_years = data |> group\_by(Year) |> summarise(Totalsales= sum(Global\_Sales))|> arrange(desc(Totalsales))

sales\_years

###maximum sales was for 2008 released games.

###lets put this in graph

ggplot(sales\_years, aes(x= Year, y = Totalsales))+ geom\_col(fill = 'red')+ labs(title= 'game released Year by sales', x= 'Year', y= 'Total Sales')

#####lets explore more on sales distribution w.r.t area

colnames(data)

area\_sales= data |> select(NA\_Sales, EU\_Sales, JP\_Sales, Other\_Sales, Genre) |> group\_by(Genre) |> pivot\_longer(c("NA\_Sales","EU\_Sales","JP\_Sales","Other\_Sales" ), values\_to = "sales")

area\_sales

ggplot(area\_sales, aes(x= Genre, y = sales, fill= name))+geom\_col()+ labs(title = "Genre Sales per Region", x= "Genre", y = "Sales")

#####Through graph we can see the genre sales per region

###lets now calculate the region sales w.r.t years

colnames(data)

Region\_sales\_year= data |> select (Year, NA\_Sales, EU\_Sales, JP\_Sales, Other\_Sales, Global\_Sales) |> group\_by(Year) |> summarise(EUsales = sum(EU\_Sales), JPsales = sum(JP\_Sales), NAsales = sum(NA\_Sales), Others = sum(Other\_Sales), Global= sum(Global\_Sales) )

Region\_sales\_year

ggplot(Region\_sales\_year) + geom\_line(aes(x= Year, y= EUsales, colour = "EUsales"), size=1)+ geom\_line(aes(x= Year, y= JPsales, colour = "JPsales"), size=1)+ geom\_line(aes(x= Year, y= NAsales, colour = "NAsales"), size=1)+ geom\_line(aes(x= Year, y= Others, colour = "Others"), size=1)+ labs(title = "Sales tred", y = "sales")

####We can clearly see through graph sales in NA region is the maximum

###lets find out which are the top 10 games

Name\_sales = data |> group\_by(Name) |> summarise(Totalsales= sum(Global\_Sales))|> arrange(desc(Totalsales))

head(Name\_sales, 10)

######Lets find out top 10 publisher according to global sales

Publisher\_sales = data |> group\_by(Publisher) |> summarise(Totalsales= sum(Global\_Sales))|> arrange(desc(Totalsales))

head(Publisher\_sales, 10)

######Lets find out top 10 Platform according to global sales

Platform\_sales = data |> group\_by(Platform) |> summarise(Totalsales= sum(Global\_Sales))|> arrange(desc(Totalsales))

head(Platform\_sales, 10)

#######CONCLUSIONS

#1. Action games are more popular than other genres

#2. PS2 platform had the highest sales globally, followed by X360

#3. The year 2008 had the highest number of games released

#6. The sales trend shows that NA region had the highest sales.

#5. Nintendo had the highest sales globally

##########THANK YOU

print("Thank You")