Steam games 2024

Correlation and hypothesis testing analysis

Made by:Hampus Nordholm

2024-09-15

Intro

This analysis explores a dataset of the top 1500 games on Steam by revenue from (January 1 2024) to (September 9 2024). Exploratory data analysis (EDA) was performed using ggplot2 to understand relationships among review scores, prices, and copies sold. Additionally,correlation analysis and also hypothesis testing was conducted to determine whether games released in the first quarter tend to have higher sales compared to those released in the second quarter. This document includes the processes of data cleansing, wrangling, visualization, and statistical testing to derive insights from raw data.

Solution summary

The correlation analysis revealed that games with high sales volumes (bin: 38,313 and above) are positively correlated with price ranges above $19.99 USD, suggesting that less expensive games are less popular on Steam. Additionally, games published by AA and AAA publishers show a positive correlation with high sales volumes, whereas Indie publishers exhibit a negative correlation. The highest average sales are observed in April, February, and March. Non-parametric hypothesis testing, conducted due to non-normal distribution of the sales data, found no significant evidence that games released in the first quarter have higher sales compared to those in the second quarter, especially considering that the full year's sales data is incomplete.

Core analysis

```
#LIBRARIES


#Data analysis
library(tidyverse)
library(correlationfunnel)
library(skimr)
library(lubridate)
```

```
#DATA IMPORT

steam_2024_tbl <- read_csv("Steam_2024_bestRevenue_1500.csv")
```

```
#DATA EXAMINATION

steam_2024_tbl %>% skim()
```

Table 1: Data summary

Name	Piped data
Number of rows	1500
Number of columns	11
Column type frequency:	
character	5
numeric	6
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
name	0	1	3	207	0	1500	0
releaseDate	0	1	10	10	0	235	0
publisherClass	0	1	2	8	0	4	0
publishers	1	1	1	60	0	1131	0
developers	2	1	2	70	0	1406	0

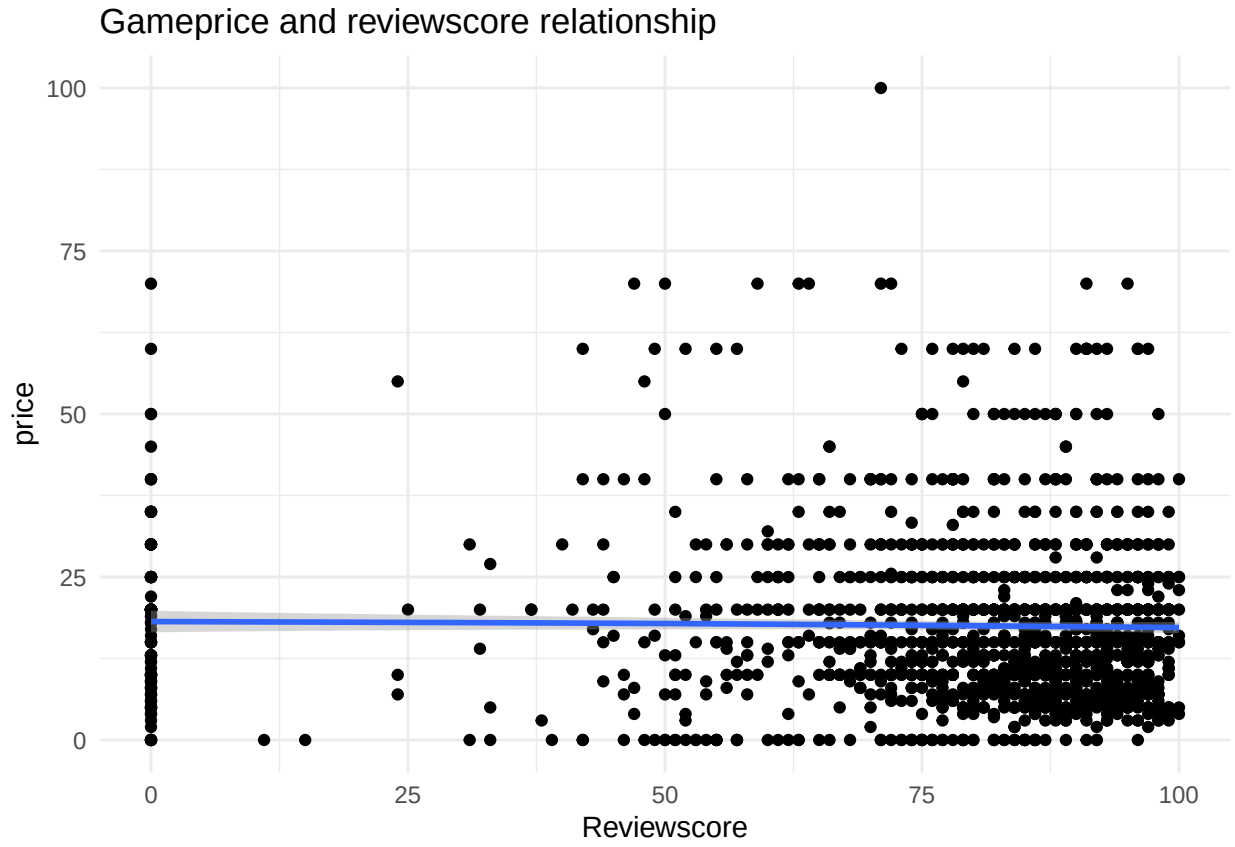
Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
copiesSold	0	1	141482.57	1132756.66	593	4918.75	11928.50	37869.75	30739148.00	
price	0	1	17.52	12.65	0	9.99	14.99	19.99	99.99	
revenue	0	1	2632381.98	27810239.62	20674	45504.25	109053.00	455156.75	837793356.00	
avgPlaytime	0	1	12.56	21.54	0	3.56	6.76	13.10	296.33	
reviewScore	0	1	76.20	24.32	0	72.00	83.00	92.00	100.00	
steamId	0	1	2183788.46	606772.46	24880	1792795.00	2321985.00	2693227.50	3107330.00	

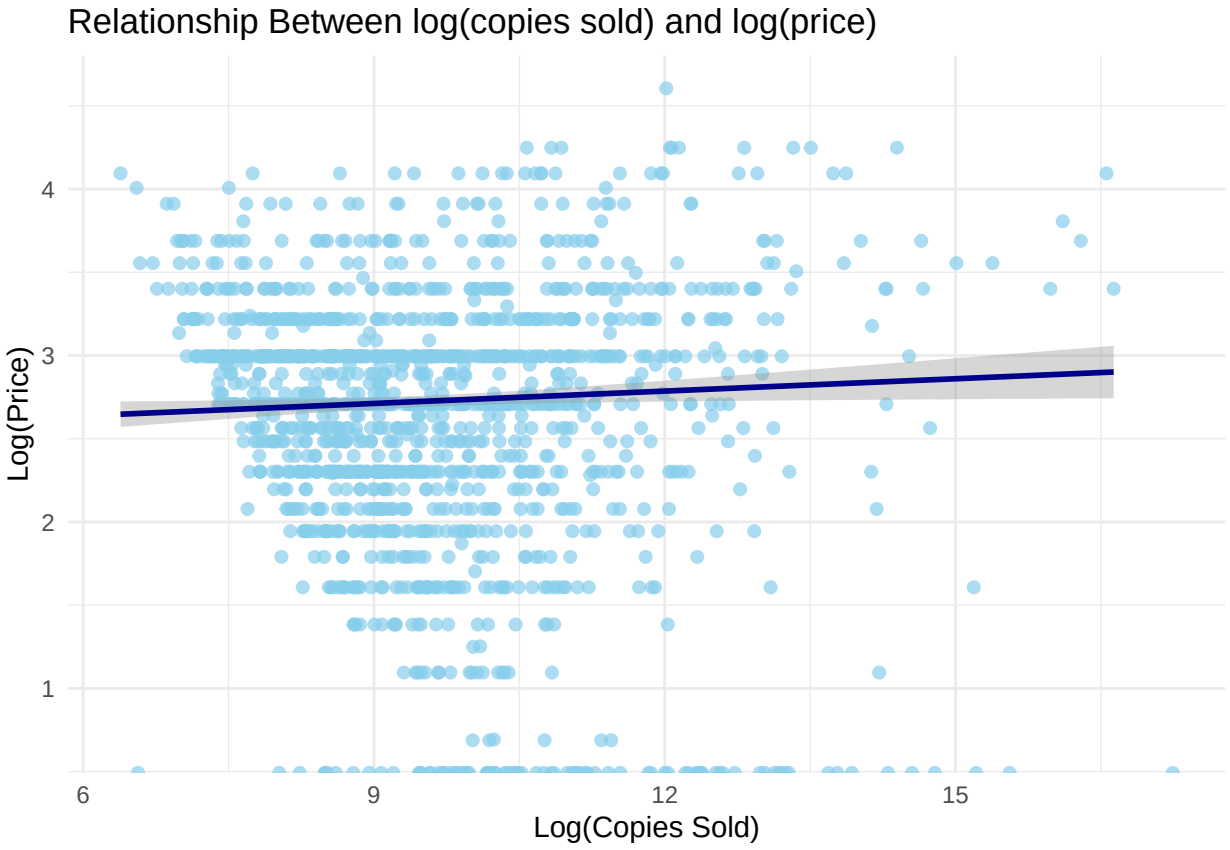
```
steam_2024_tbl %>% View()
```

```
# NA removal -- 3 OBS.

steam_2024_tbl <- steam_2024_tbl %>% drop_na()

#Feature enginering -- date -> monthcolumn

steam_2024_tbl <- steam_2024_tbl %>%
  mutate(releasemonth = month(dmy(releaseDate),label=TRUE,abbr=TRUE)) %>%
  mutate(releasemonth=as.factor(releasemonth))
```

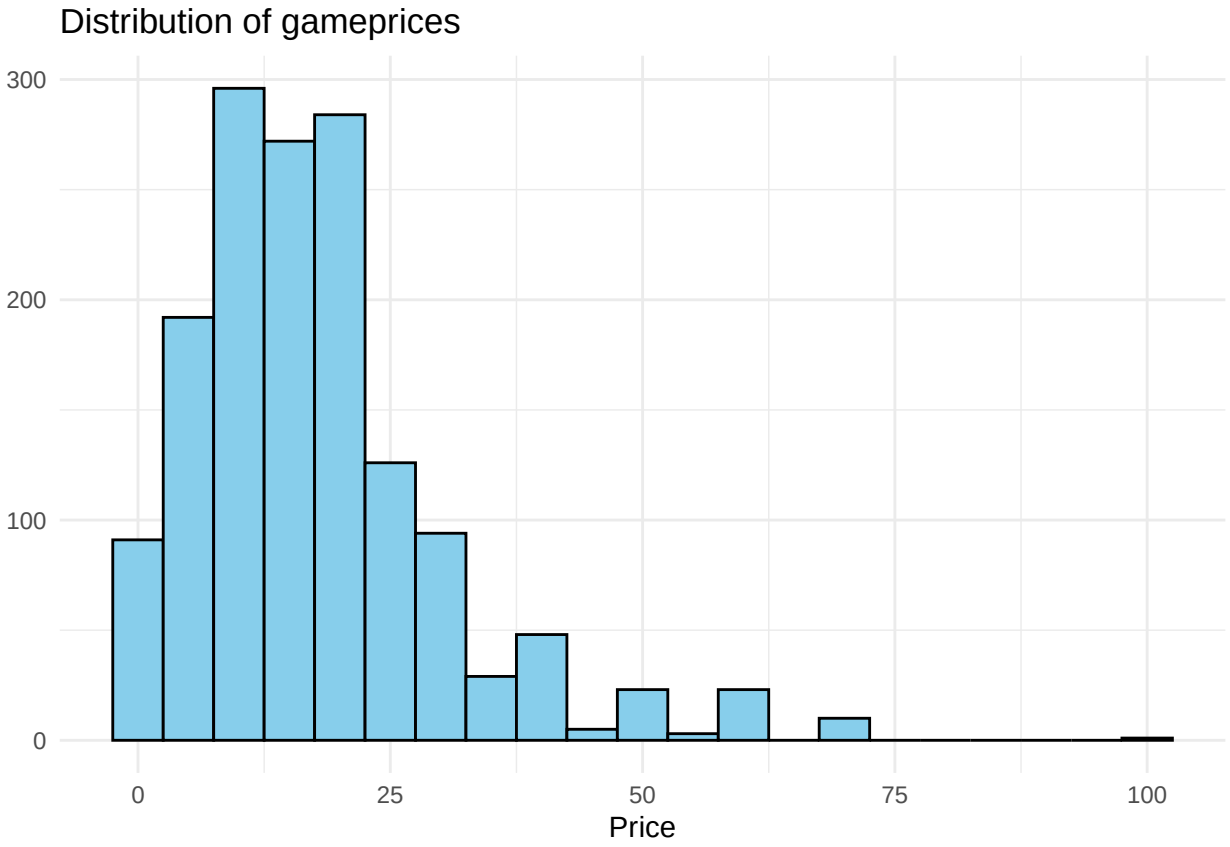
```r
# EXPLORATORY DATA ANALYSIS (GGPLOT) --

# Reviewscore - price | relationship

steam_2024_tbl %>%
  ggplot(aes(reviewScore,price))+
  geom_point(size=1.5,color="black")+
  geom_smooth()+
  theme_minimal()+
  labs(title="Gameprice and reviewscore relationship",x="Reviewscore")
```



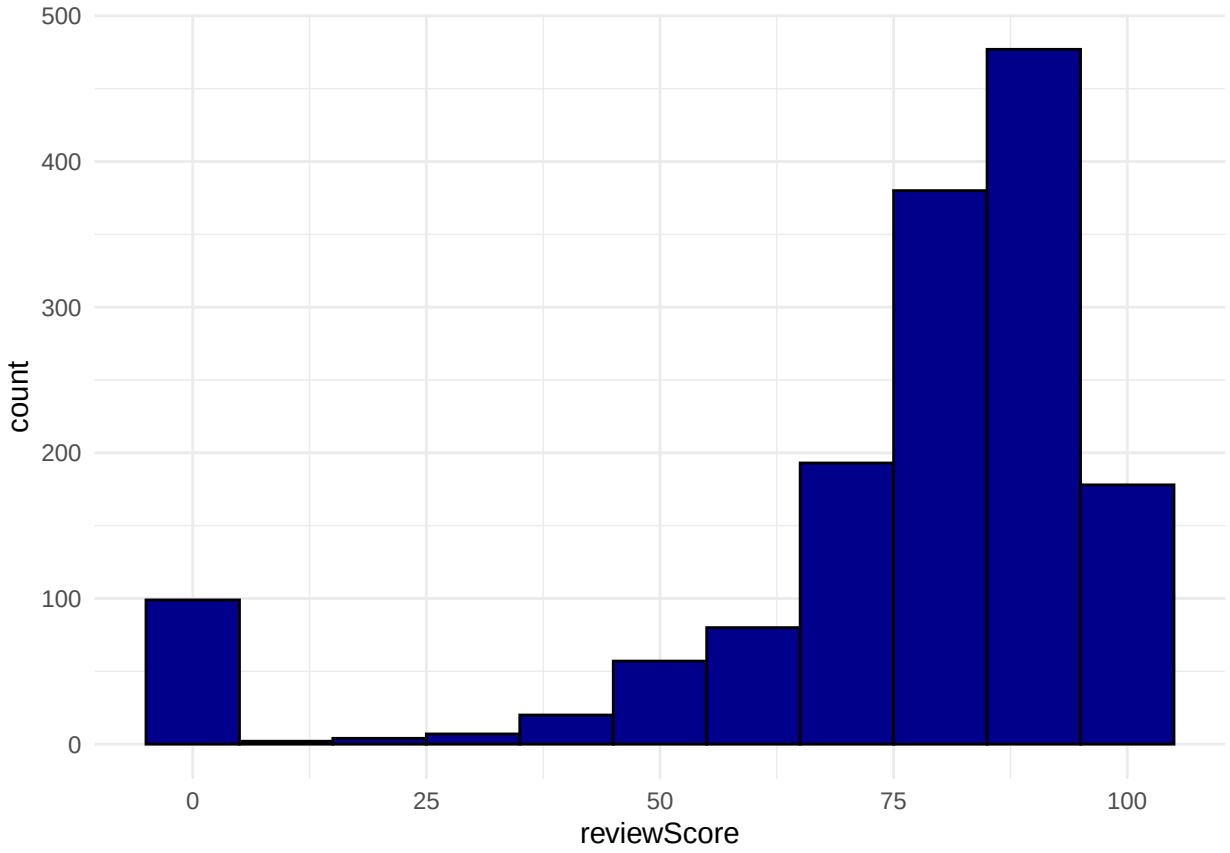

```r
# Log(copiessold) - log(price) | relationship

steam_2024_tbl %>%
  ggplot(aes(x = log(copiesSold), y = log(price))) +
  geom_point(size = 1.8, color = "skyblue", alpha = 0.7) +
  geom_smooth(method = "lm", color = "darkblue") +
  labs(
  x = "Log(Copies Sold)",
  y = "Log(Price)",
  title = "Relationship Between log(copies sold) and log(price)")+
  theme_minimal()
```



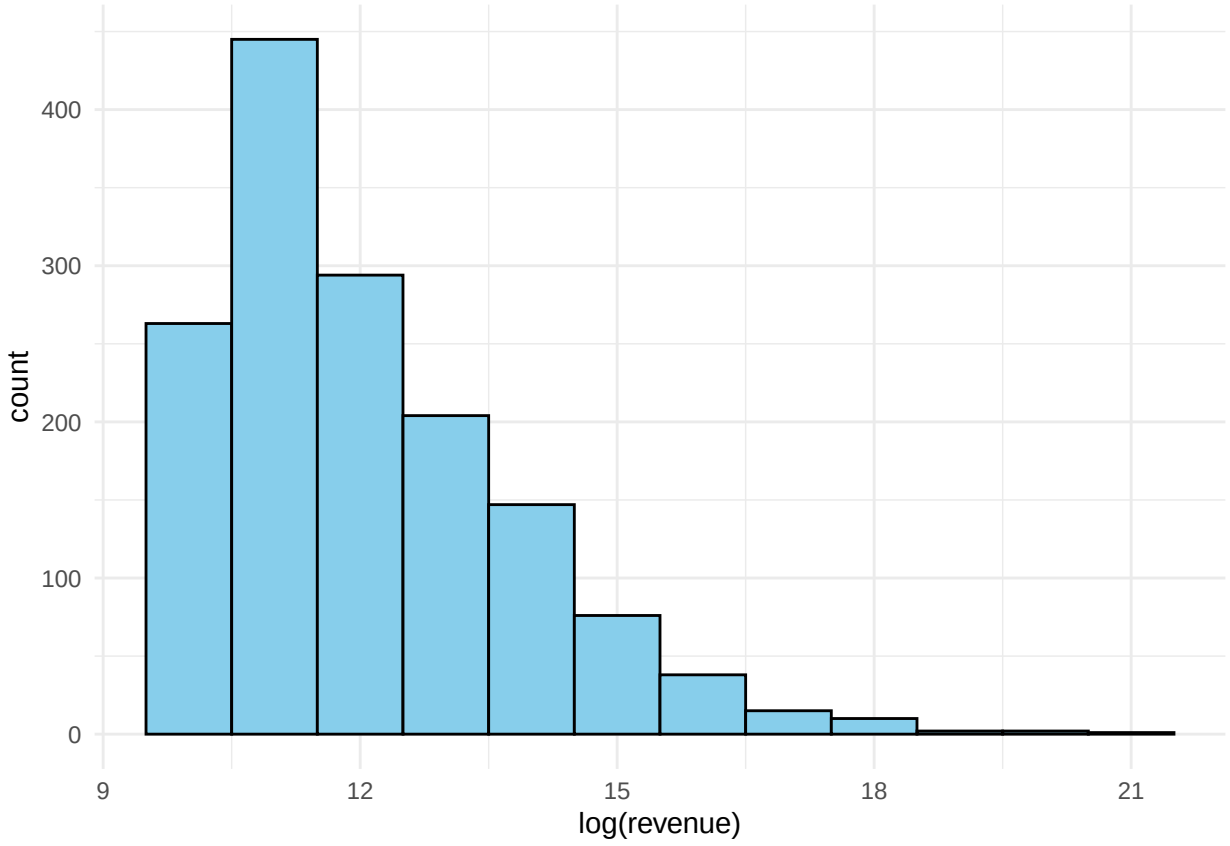

```
# PRICE HISTOGRAM --

steam_2024_tbl %>%
  ggplot(aes(price))+
  geom_histogram(binwidth =5,fill="skyblue",color="black")+
  theme_minimal()+
  labs(title="Distribution of gameprices",x="Price",y=NULL)
```



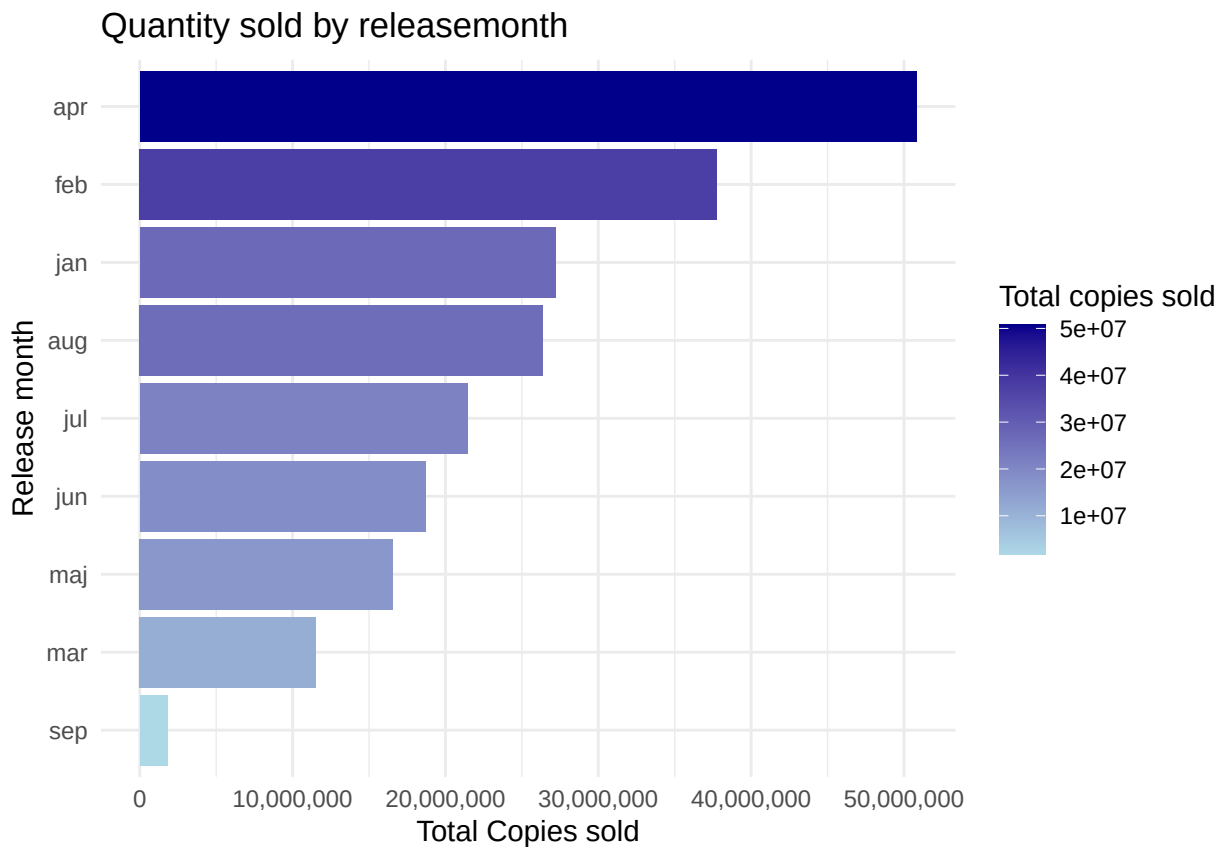

```
# REVIEWSCORE HISTOGRAM --

steam_2024_tbl %>%
  ggplot(aes(reviewScore))+
  geom_histogram(binwidth=10,fill="darkblue",color="black")+
  theme_minimal()
```



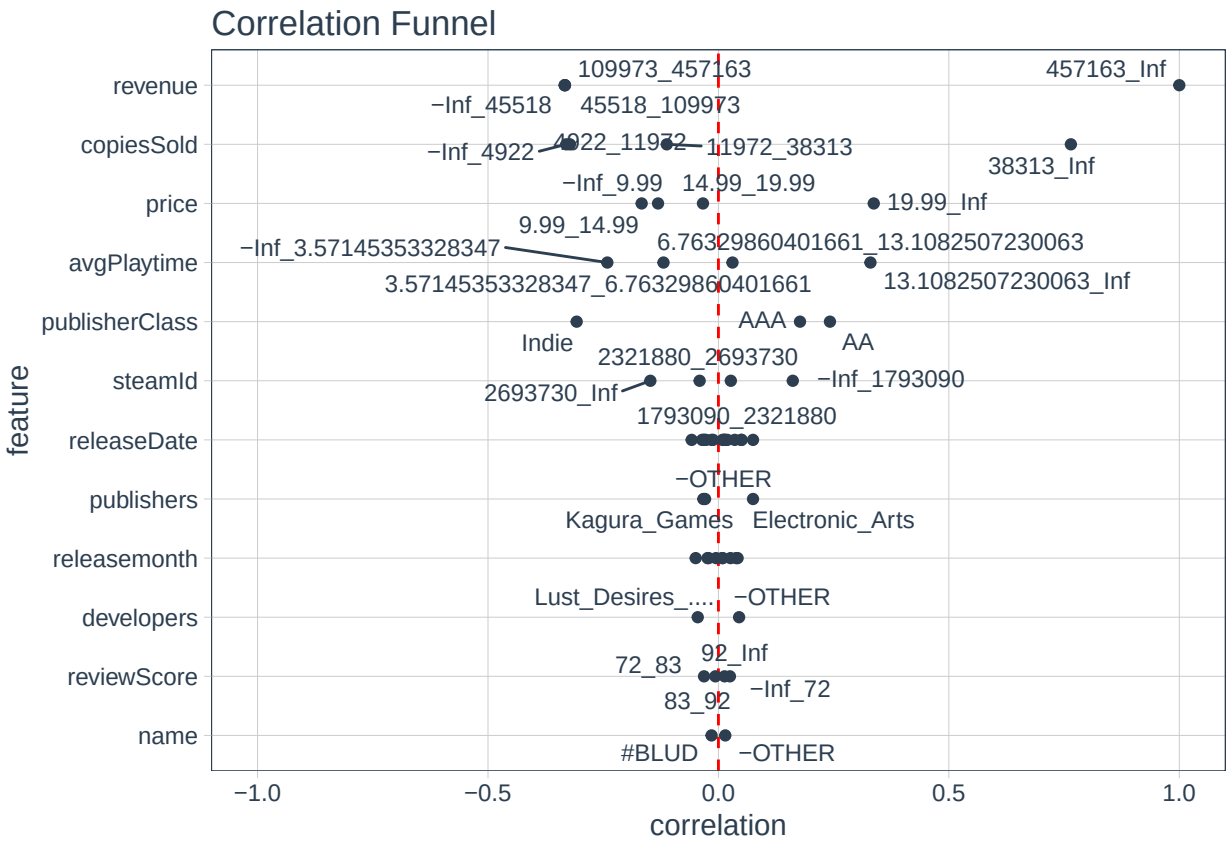

```
# REVENUE HISTOGRAM --

steam_2024_tbl %>%
  ggplot(aes(log(revenue)))+
  geom_histogram(binwidth =1,fill="skyblue",color="black")+
  theme_minimal()
```



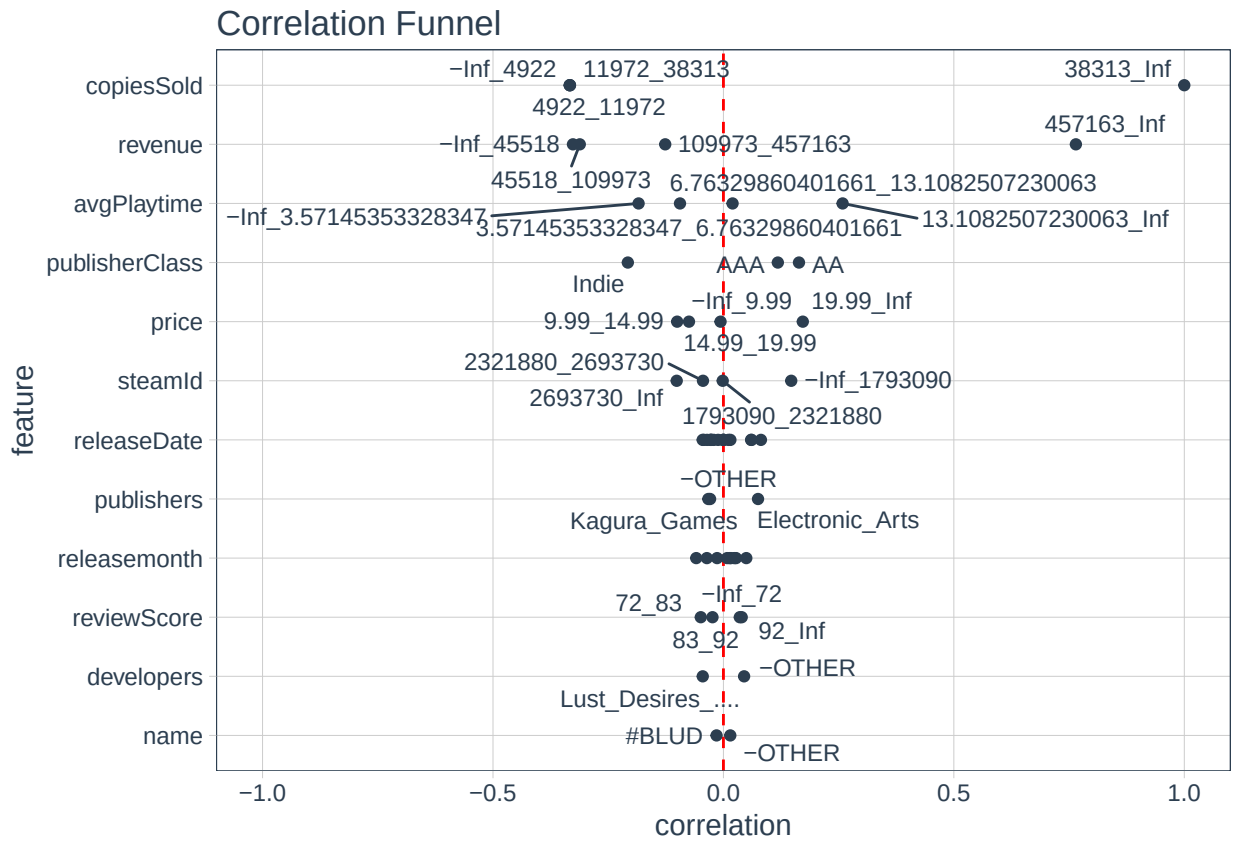

```
# Total Q sold by releasemonth

steam_2024_tbl %>%
  group_by(releasemonth) %>%
  summarise(total_q_sold = sum(copiesSold)) %>%
  mutate(releasemonth=fct_reorder(releasemonth, total_q_sold)) %>%
  ggplot(aes(x = total_q_sold, y = releasemonth, fill = total_q_sold))+
  geom_col()+
  scale_fill_gradient(low="lightblue",high="darkblue")+
  labs(x="Total Copies sold",y="Release month",fill ="Total copies sold",
  title ="Quantity sold by releasemonth")+
  theme_minimal()+
  scale_x_continuous(labels=scales::label_comma())
```




```
# CORRELATION ANALYSIS

steam_2024_tbl %>%
  binarize() %>%
  correlate(revenue__457163_Inf) %>%
  plot_correlation_funnel()
```



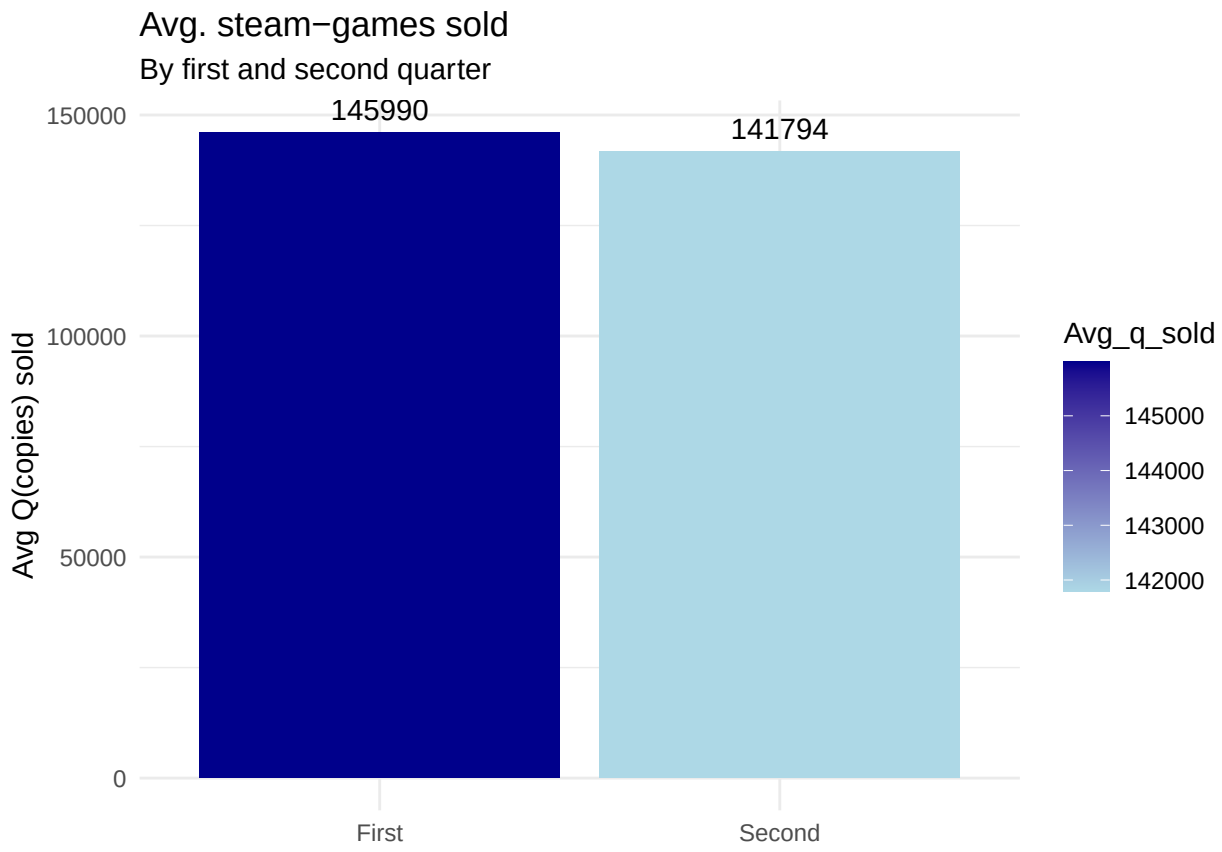

```
steam_2024_tbl %>%
  binarize() %>%
  correlate(copiesSold__38313_Inf) %>%
  plot_correlation_funnel()
```




```
# -- Hypothesis testing --

# Hypothesis: does games released during first quarter tend to be sold more then
# games released in second quarter? --

# New feature: first or. second quarter release --

steam_quarter_tbl <- steam_2024_tbl %>%
  filter(releasemonth %in% c("jan","feb","mar","apr","maj","jun")) %>%
  mutate(Quarter=case_when(releasemonth %in% c("jan","feb","mar")~"First",
                           releasemonth %in% c("apr","maj","jun")~"Second"))

# Avg. copies sold first vs. second quarter --

steam_quarter_tbl %>%
  group_by(Quarter) %>%
  summarise(Avg_q_sold=mean(copiesSold))
```

```
## # A tibble: 2 x 2
##   Quarter Avg_q_sold
##   <chr>        <dbl>
## 1 First      145990.
## 2 Second     141794.
```

```
# GGPLOT VISUAL --

steam_quarter_tbl %>%
  group_by(Quarter) %>%
  summarise(Avg_q_sold = mean(copiesSold)) %>%
  ggplot(aes(x =Quarter,y=Avg_q_sold,fill=Avg_q_sold))+
  geom_col()+
  scale_fill_gradient(low="lightblue",high="darkblue")+
  theme_minimal()+
  labs(title="Avg. steam-games sold",subtitle="By first and second quarter",
       x=NULL,y="Avg Q(copies) sold")+
  geom_text(aes(label=round(Avg_q_sold,0)),vjust=-0.6)
```




```
# Testing for normal distribution by quarter 1  --

steam_quarter_tbl %>%
  filter(Quarter=="First") %>%
  pull(copiesSold) %>%
  shapiro.test()
```

```
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.10963, p-value < 2.2e-16
```

```r
# First quarter not normal distributed -> non-parametric test

# Wilcox hypothesis testing --

wilcox.test(copiesSold~Quarter,data=steam_quarter_tbl)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  copiesSold by Quarter
## W = 164552, p-value = 0.3138
## alternative hypothesis: true location shift is not equal to 0
```