Project

Ian Ou, Miller Kodish, Vinay Pundith

2025-11-18

```
library(here)
```

```
## Warning: package 'here' was built under R version 4.5.2
```

```
## here() starts at C:/Users/ianou/OneDrive/Desktop/STAT-527-Final-Project-main
```

```
library(ggplot2)
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 4.5.2
```

```
## randomForest 4.7-1.2
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

Load data and merge the two csv files from Kaggle

```
gpu_benchmarks <- read.csv(here("Datasets", "Kaggle", "GPU_benchmarks_v7.csv"))
gpu_scores <- read.csv(here("Datasets", "Kaggle", "GPU_scores_graphicsAPIs.csv"))
gpu_benchmarks$gpu_name <- tolower(trimws(gpu_benchmarks$gpuName))
gpu_scores$gpu_name     <- tolower(trimws(gpu_scores$Device))

# merge both kaggle csv files
merged_gpu <- merge(
  gpu_benchmarks,
  gpu_scores,
  by = "gpu_name"
)

cat("Rows in PassMark dataset :", nrow(gpu_benchmarks), "\n")
```

```
## Rows in PassMark dataset : 2317
```

```
cat("Rows in Geekbench dataset:", nrow(gpu_scores), "\n")
```

```
## Rows in Geekbench dataset: 1213
```

```
cat("Rows in merged dataset   :", nrow(merged_gpu), "\n\n")
```

```
## Rows in merged dataset   : 647
```

```r
merged_gpu$gpuName <- NULL
merged_gpu$Device <- NULL
head(merged_gpu)
```

```
##        gpu_name G3Dmark G2Dmark price gpuValue TDP powerPerformance testDate
## 1       a40-12q    5573     198    NA       NA  NA               NA     2022
## 2 firepro m4000    1597     410 72.83    21.92  NA               NA     2012
## 3 firepro m4100    1059     623    NA       NA  NA               NA     2015
## 4 firepro m4150     999     207    NA       NA  NA               NA     2015
## 5 firepro m4170    1067     290    NA       NA  NA               NA     2015
## 6 firepro m5100    2103     800    NA       NA  NA               NA     2014
##      category Manufacturer  CUDA Metal OpenCL Vulkan
## 1     Unknown       Nvidia 95329    NA 156643     NA
## 2 Workstation          AMD    NA    NA   6494     NA
## 3 Workstation          AMD    NA    NA   5067     NA
## 4     Unknown          AMD    NA    NA   5063   6685
## 5     Unknown          AMD    NA    NA   6347     NA
## 6 Workstation          AMD    NA    NA   9305  10692
```

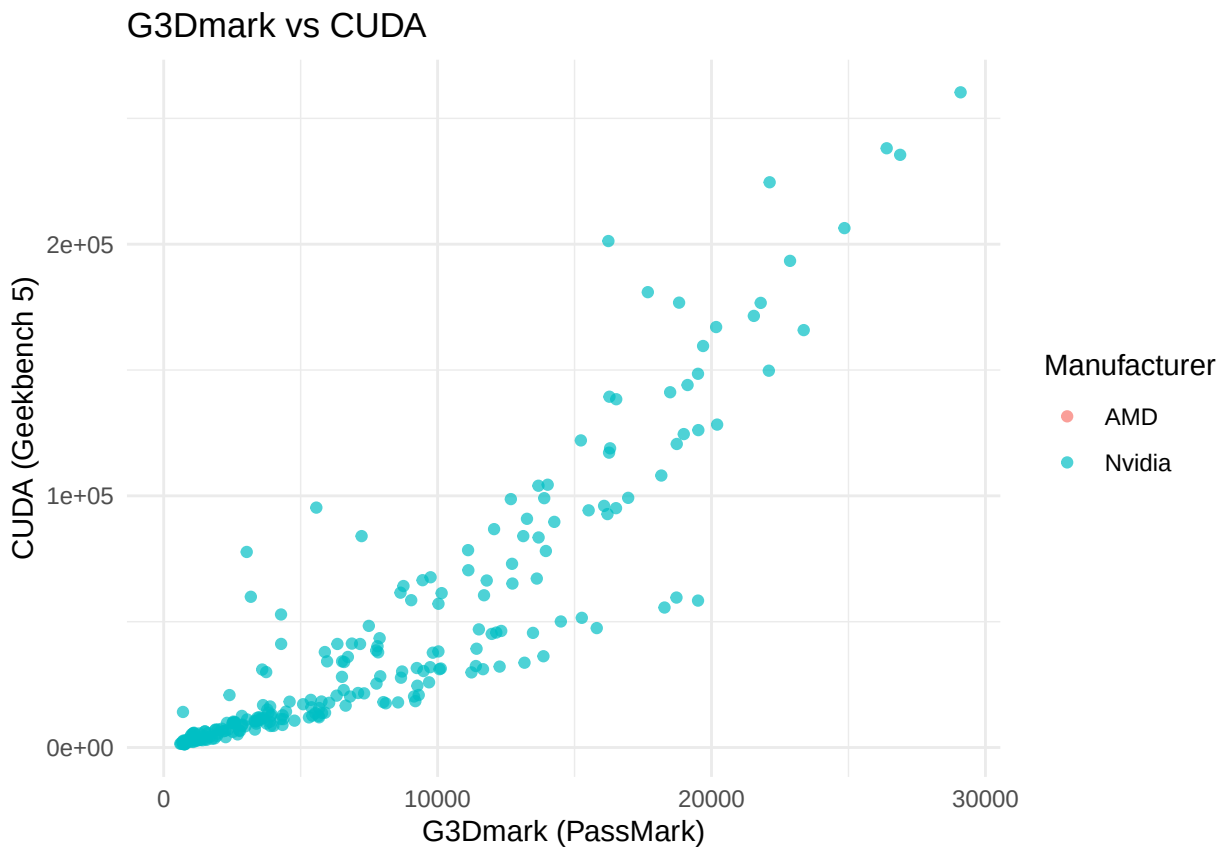
Generate plot to see relationship between CUDA/OpenCL/Vulkan to G3dmark

```r
plot_scatter <- function(x_col, y_col, data) {
  ggplot(data, aes_string(x = x_col, y = y_col, color = "Manufacturer")) +
    geom_point(alpha = 0.7) +
    theme_minimal() +
    labs(
      title = paste(x_col, "vs", y_col),
      x = paste(x_col, "(PassMark)"),
      y = paste(y_col, "(Geekbench 5)"),
      color = "Manufacturer"
    )
}

# G3Dmark vs CUDA
print(plot_scatter("G3Dmark", "CUDA", merged_gpu))
```

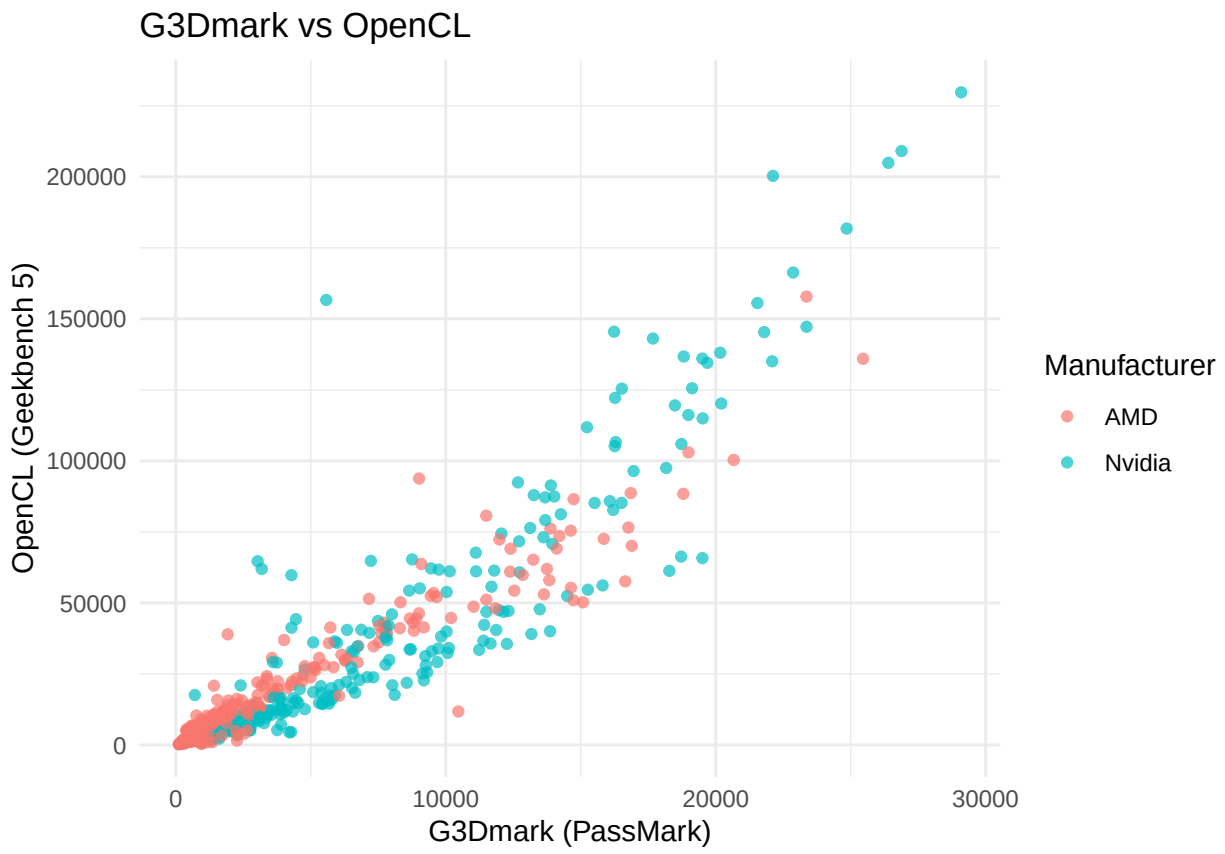
```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with `aes()`.
## i See also `vignette("ggplot2-in-packages")` for more information.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## Warning: Removed 419 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



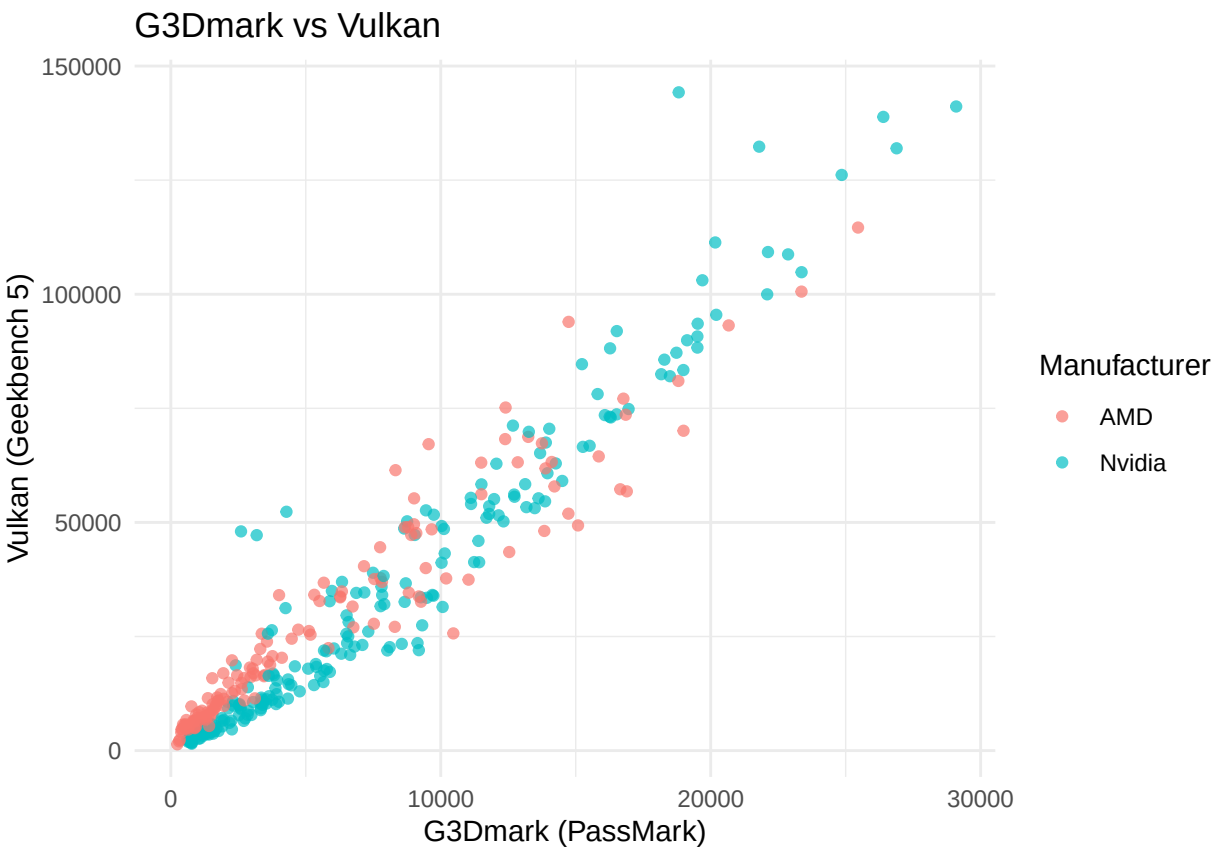

```
# G3Dmark vs OpenCL
print(plot_scatter("G3Dmark", "OpenCL", merged_gpu))
```

```
## Warning: Removed 11 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



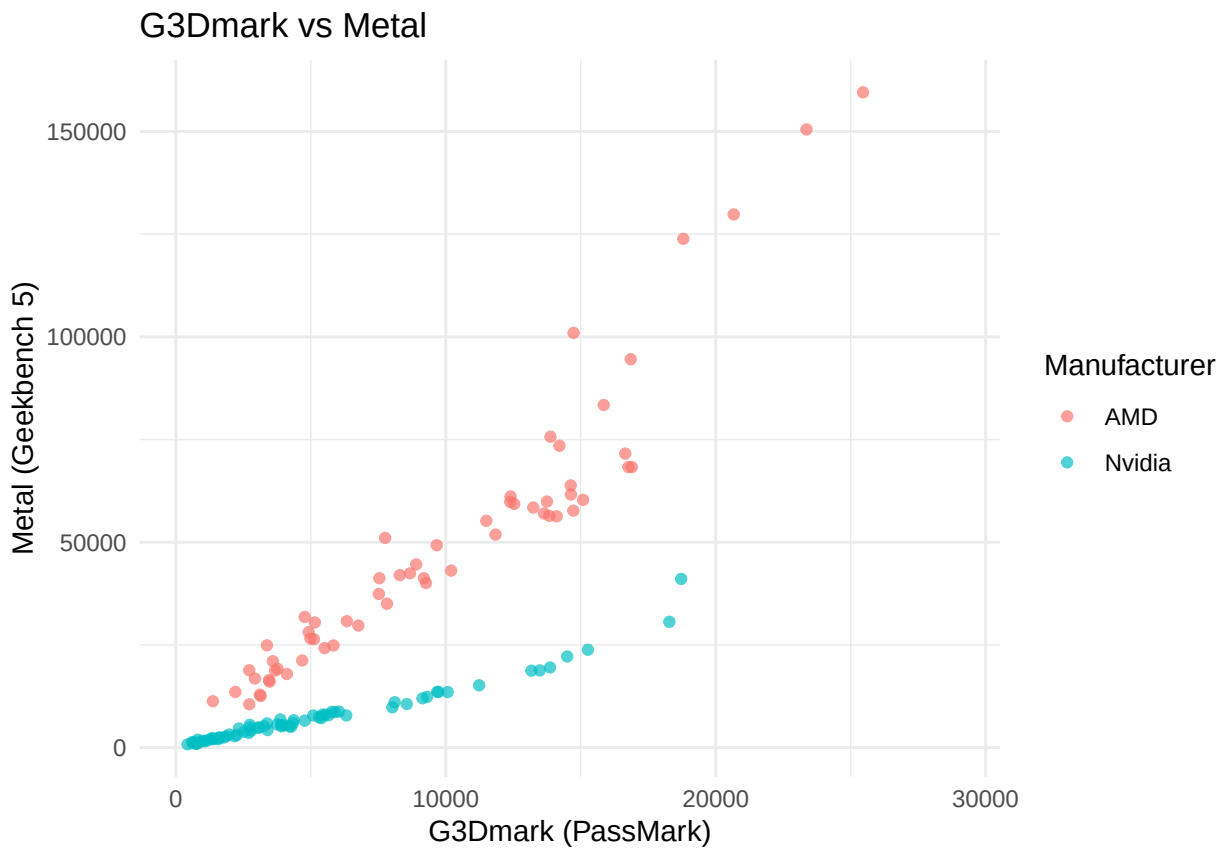

```
# G3Dmark vs Vulkan
print(plot_scatter("G3Dmark", "Vulkan", merged_gpu))
```

```
## Warning: Removed 298 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

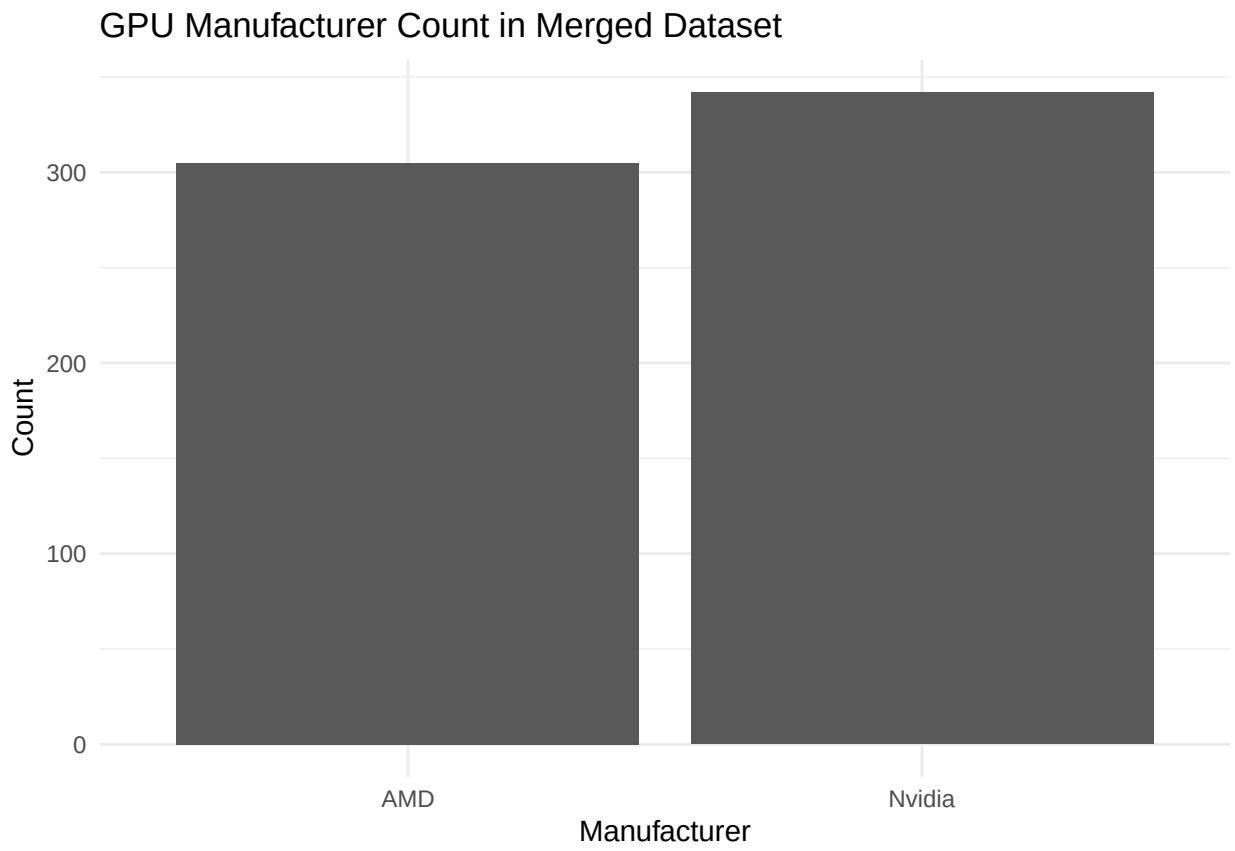



```
# G3Dmark vs Metal (mostly Apple GPUs)
if ("Metal" %in% names(merged_gpu)) {
  print(plot_scatter("G3Dmark", "Metal", merged_gpu))
}
```

```
## Warning: Removed 514 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



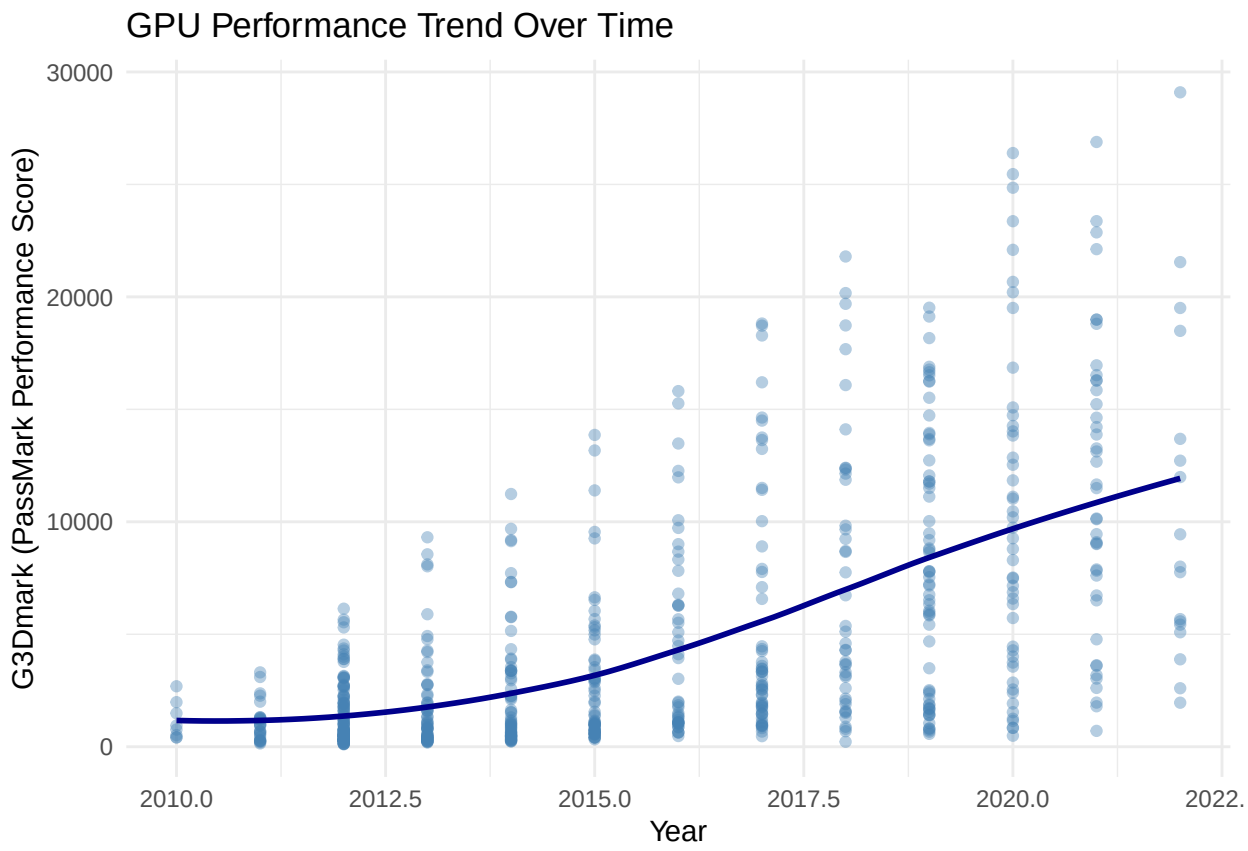

```r
# Manufacturer Distribution
ggplot(merged_gpu, aes(x = Manufacturer)) +
  geom_bar() +
  theme_minimal() +
  labs(
    title = "GPU Manufacturer Count in Merged Dataset",
    x = "Manufacturer",
    y = "Count"
  )
```



Plot the trends

```
# gpu performance over time
ggplot(merged_gpu, aes(x = testDate, y = G3Dmark)) +
  geom_point(alpha = 0.4, color = "steelblue") +
  geom_smooth(method = "loess", se = FALSE, color = "darkblue") +
  theme_minimal() +
  labs(
    title = "GPU Performance Trend Over Time",
    x = "Year",
    y = "G3Dmark (PassMark Performance Score)"
  )
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

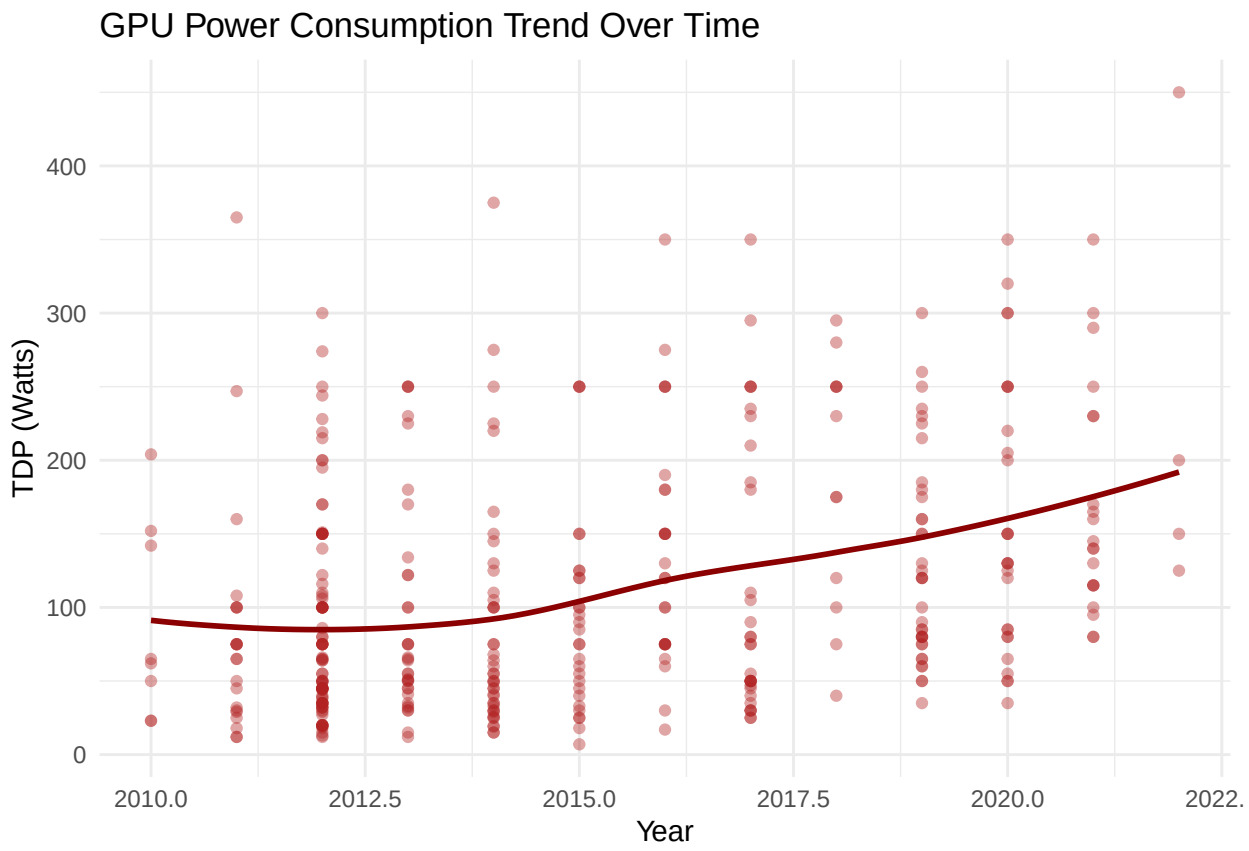



```
# tdp over time
ggplot(merged_gpu, aes(x = testDate, y = TDP)) +
  geom_point(alpha = 0.4, color = "firebrick") +
  geom_smooth(method = "loess", se = FALSE, color = "darkred") +
  theme_minimal() +
  labs(
    title = "GPU Power Consumption Trend Over Time",
    x = "Year",
    y = "TDP (Watts)"
  )
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 266 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

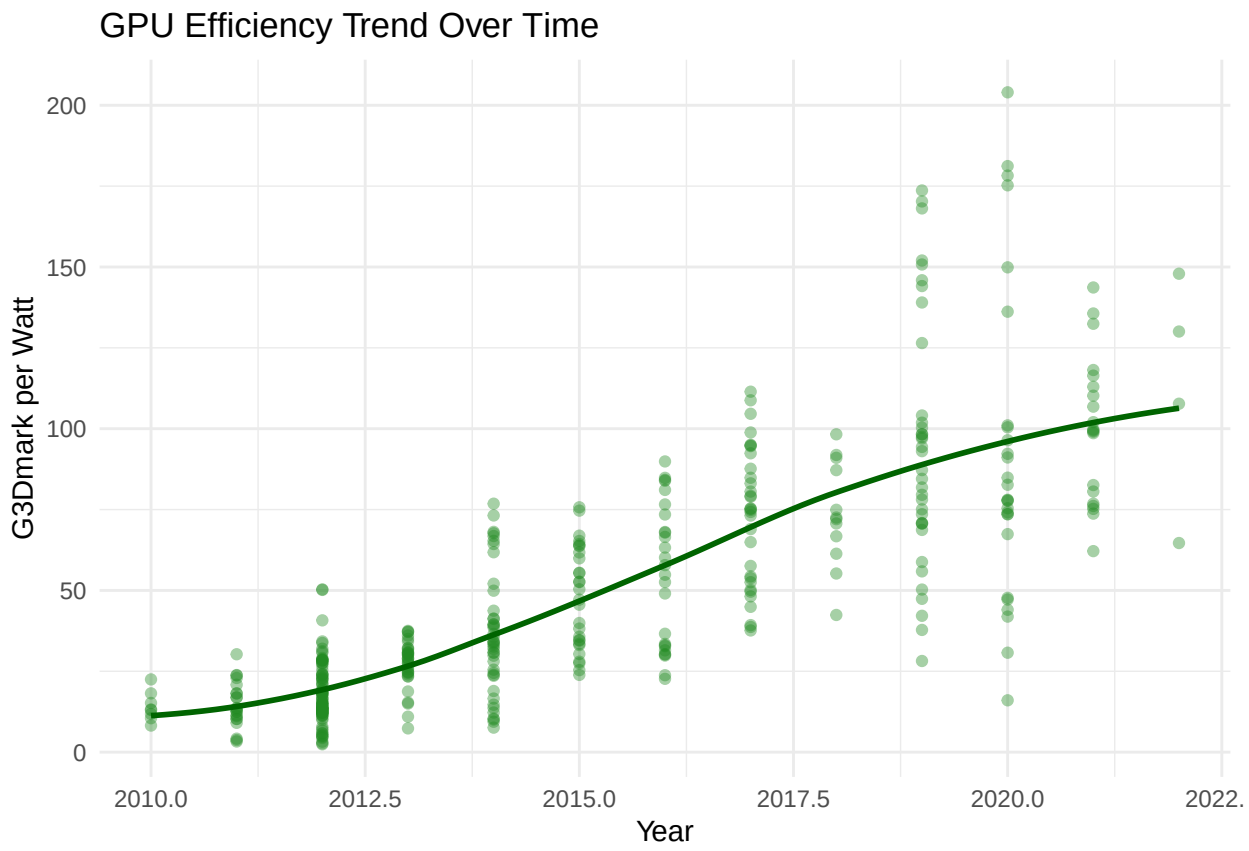
```
## Warning: Removed 266 rows containing missing values or values outside the scale range
## (`geom_point()`).
```




```
# performance per watt over time
merged_gpu$PerfPerWatt <- merged_gpu$G3Dmark / merged_gpu$TDP
ggplot(merged_gpu, aes(x = testDate, y = PerfPerWatt)) +
  geom_point(alpha = 0.4, color = "forestgreen") +
  geom_smooth(method = "loess", se = FALSE, color = "darkgreen") +
  theme_minimal() +
  labs(
    title = "GPU Efficiency Trend Over Time",
    x = "Year",
    y = "G3Dmark per Watt"
  )
```

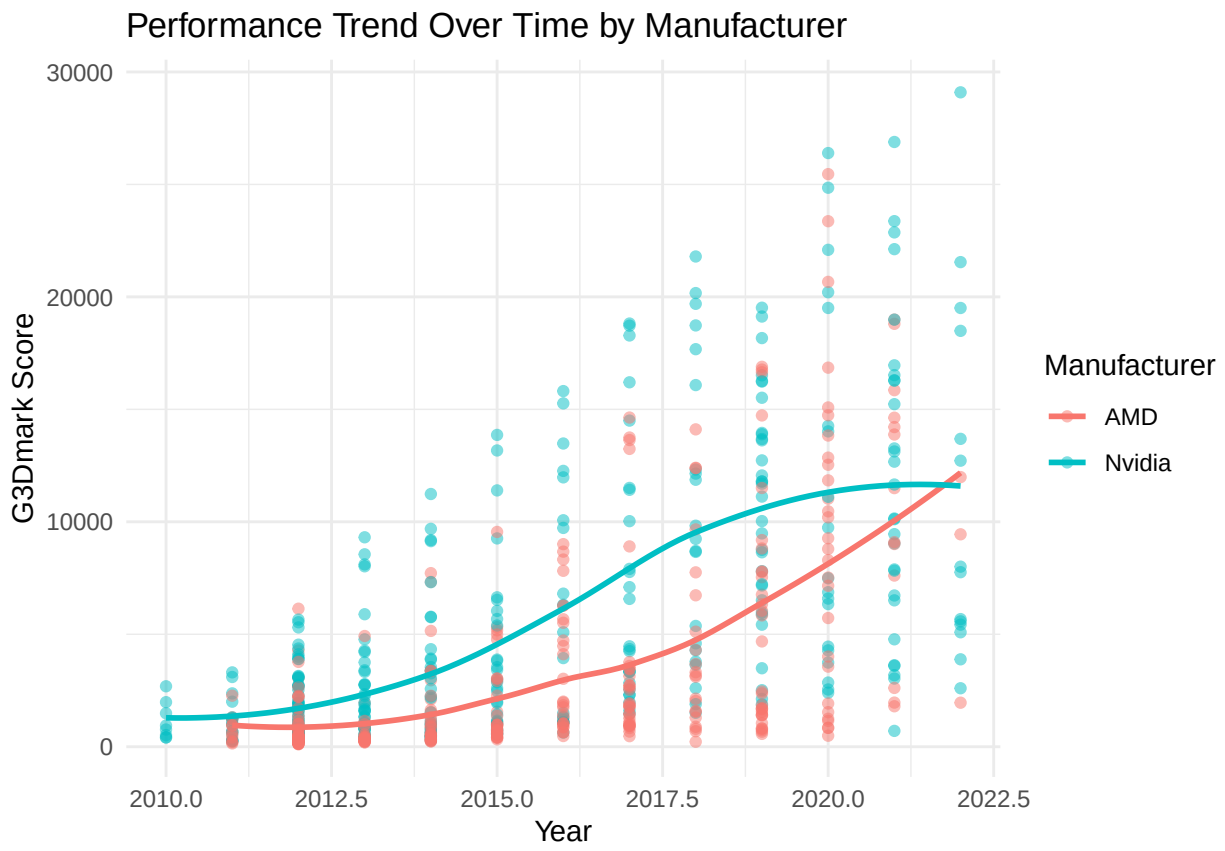
```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 266 rows containing non-finite outside the scale range
## (`stat_smooth()`).
## Removed 266 rows containing missing values or values outside the scale range
## (`geom_point()`).
```




```
# amd vs nvidia
ggplot(merged_gpu, aes(x = testDate, y = G3Dmark, color = Manufacturer)) +
  geom_point(alpha = 0.5) +
  geom_smooth(se = FALSE) +
  theme_minimal() +
  labs(
    title = "Performance Trend Over Time by Manufacturer",
    x = "Year",
    y = "G3Dmark Score"
  )
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



Build a simple linear regression model

```
# build a linear regression model
merged_gpu$Manufacturer <- as.factor(merged_gpu$Manufacturer)
merged_gpu$category <- as.factor(merged_gpu$category)
linear_model <- lm(G3Dmark ~ testDate + TDP + price + Manufacturer + category, data = merged_gpu)
summary(linear_model)
```

```
##
## Call:
## lm(formula = G3Dmark ~ testDate + TDP + price + Manufacturer +
##     category, data = merged_gpu)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -10006  -1412     95   1475   6633
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -2.632e+06  1.126e+05 -23.377  < 2e-16 ***
## testDate                      1.307e+03  5.591e+01  23.371  < 2e-16 ***
## TDP                           3.588e+01  2.194e+00  16.356  < 2e-16 ***
## price                         5.209e-01  1.987e-01   2.622  0.00931 **
## ManufacturerNvidia            1.608e+03  3.304e+02   4.866 2.07e-06 ***
## categoryMobile               -1.538e+02  5.279e+02  -0.291  0.77111
## categoryMobile, Workstation -1.020e+03  9.347e+02  -1.091  0.27647
## categoryUnknown               5.400e+02  1.420e+03   0.380  0.70401
```

```
## categoryWorkstation         -2.816e+02  3.861e+02  -0.729  0.46644
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2339 on 238 degrees of freedom
##   (400 observations deleted due to missingness)
## Multiple R-squared:  0.8851, Adjusted R-squared:  0.8812
## F-statistic: 229.1 on 8 and 238 DF,  p-value: < 2.2e-16
```

Build a random forest model with feature importance

```
rf_data <- merged_gpu[, c("G3Dmark", "testDate", "TDP", "price", "Manufacturer", "category")]
rf_data <- na.omit(rf_data)
rf_data$Manufacturer <- as.factor(rf_data$Manufacturer)
rf_data$category <- as.factor(rf_data$category)
nrow(rf_data)
```

```
## [1] 247
```

```
n <- nrow(rf_data)
ix <- sample(seq_len(n), size = floor(0.8 * n))
train_rf <- rf_data[ix, ]
test_rf  <- rf_data[-ix, ]
rf_model <- randomForest(
  G3Dmark ~ testDate + TDP + price + Manufacturer + category,
  data   = train_rf,
  ntree  = 500,
  mtry   = 3,
  importance = TRUE
)

pred_rf <- predict(rf_model, newdata = test_rf)
rmse <- sqrt(mean((pred_rf - test_rf$G3Dmark)^2))
mae  <- mean(abs(pred_rf - test_rf$G3Dmark))
r2   <- 1 - sum((pred_rf - test_rf$G3Dmark)^2) / sum((mean(train_rf$G3Dmark) - test_rf$G3Dmark)^2)

cat("RF RMSE:", rmse, "\n")
```

```
## RF RMSE: 2089.684
```

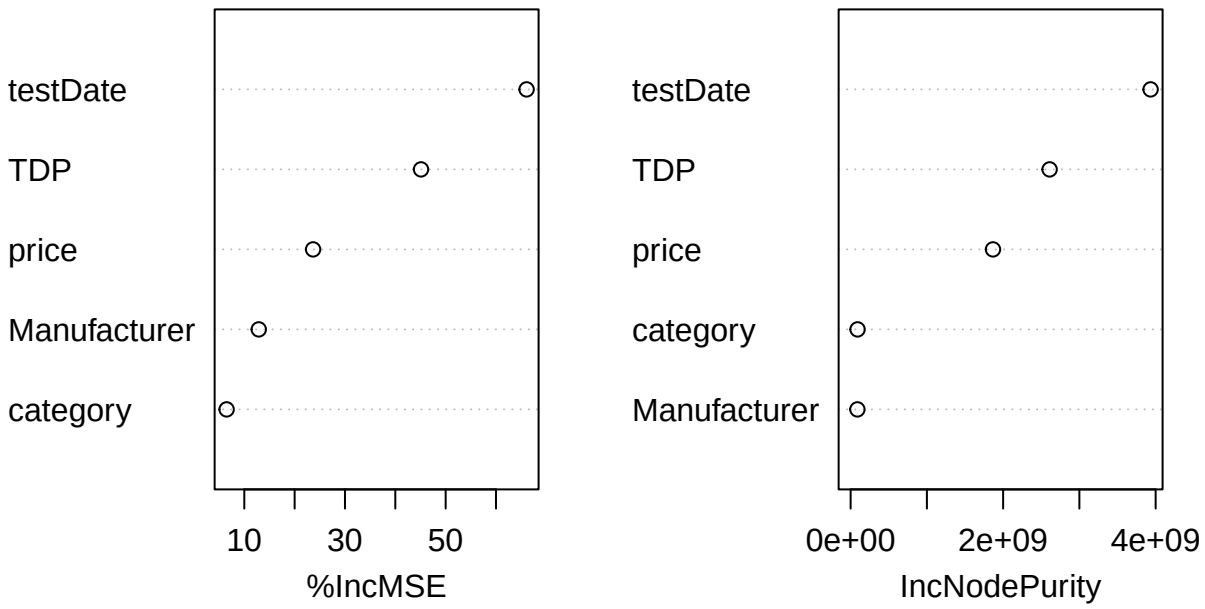
```
cat("RF MAE :", mae,  "\n")
```

```
## RF MAE : 1479.923
```

```
cat("RF R^2 :", r2,   "\n")
```

```
## RF R^2 : 0.9113455
```

```
varImpPlot(rf_model)
```


rf_model
testDate
TDP
price
Manufacturer
category
10
30
50
%IncMSE
testDate
TDP
price
category
Manufacturer
0e+00
2e+09
4e+09
IncNodePurity