Thesis Data Cleansing

Kyle Davis

3/9/2019

NCES Data Cleaning

Read in Data

Setting Path Nodes

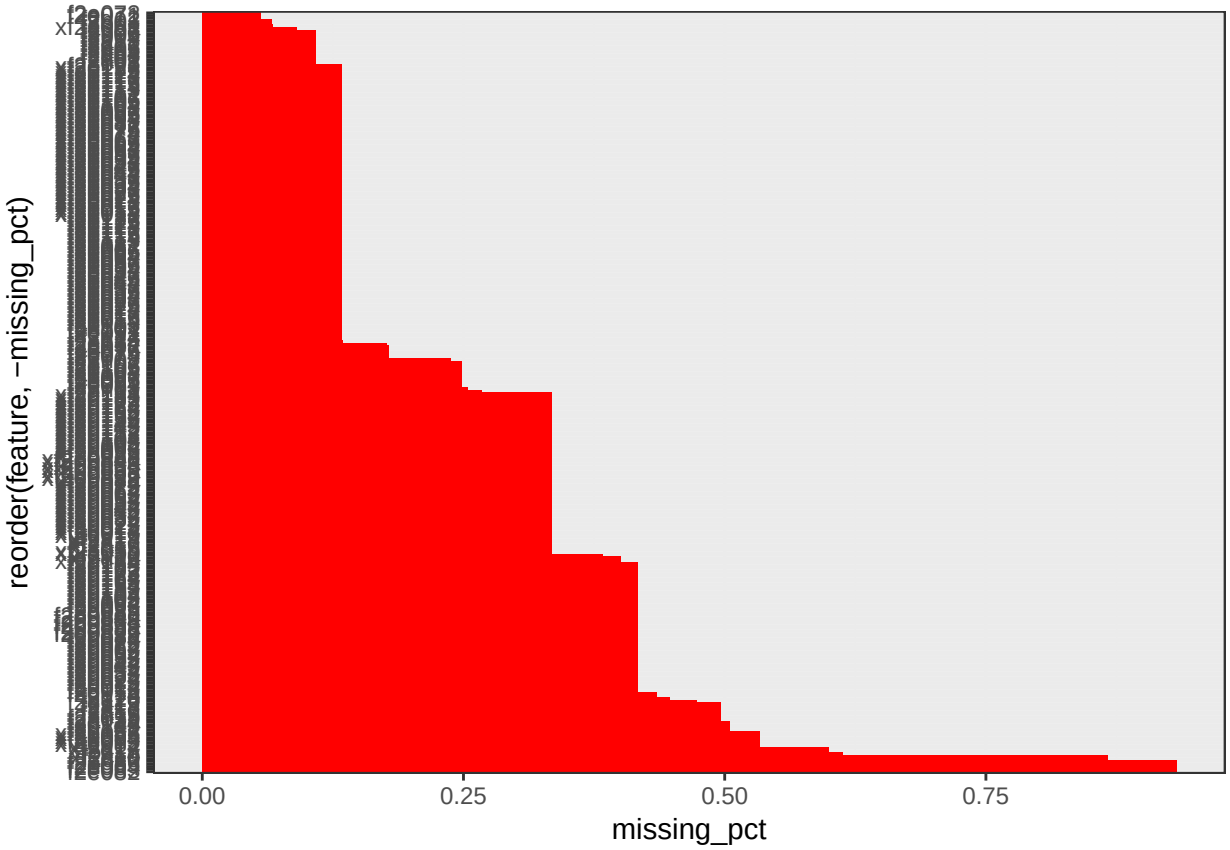
```
Finance_node <- "NCES/Finance/CSV"
Directory_node <- "NCES/Institutional Characteristic/Directory Information"
Fall_node <- "NCES/Institutional Characteristic/Fall Enrollment"
```

Read in Finance Datasets

1) Create List of all CSV files in directory
2) Initialize Dataframe by reading in first file
3) Create Year Variable to create a unique ID of UNID + year
4) Loop through all files and repeat

```
final_node <- as.character(paste(getwd(),Finance_node, sep = "/")) #location of all of the csv files

finance_files <- list.files(path = final_node) #creates a list of all the files in the specified path

path <- as.character(paste(Finance_node,finance_files[[1]], sep = "/")) #1st CSV file location

finance_data <- read.csv(path) #create the dataframe

year <- str_extract_all(finance_files[[length(finance_files)]], pattern = "[0-9]{2}") #Years are specif
year <- paste0("20",unlist(year)[[2]]) #set year equal to 20xx

finance_data <- finance_data %>%
  mutate(year = year) #create year variable in finance_dataset

colnames(finance_data) <- tolower(colnames(finance_data))

for(i in 2:length(finance_files)){
  path <- as.character(paste(Finance_node,finance_files[[i]], sep = "/"))
  placeholder <- read.csv(path) #create the dataframe

  year <- str_extract_all(finance_files[[i]], pattern = "[0-9]{2}")
  year <- paste0("20",unlist(year)[[2]])

  placeholder <- placeholder %>%
  mutate(year = year)

  colnames(placeholder) <- tolower(colnames(placeholder))
  finance_data <- finance_data %>%
    plyr::rbind.fill(placeholder)
}
```

```
missing_values <- finance_data %>% summarize_all(funs(sum(is.na(.))/n()))
missing_values <- gather(missing_values, key = "feature", value = "missing_pct")
missing_values %>%
  filter(missing_pct > 0.03) %>%
ggplot(aes(x=reorder(feature,-missing_pct), y = missing_pct)) +
geom_bar(stat="identity",fill="red")+
coord_flip()+theme_bw()
```



Selecting Relevent Variables

```
varlist_path <- paste0(getwd(),"/NCES/Finance/Finance_Varlist.xlsx")
varlist <- readxl::read_xlsx(path = varlist_path, sheet = 3)

NCES_varnames <- c(tolower(varlist$varname), 'year')

finance_data <- finance_data %>%
  dplyr::select_(.dots = NCES_varnames)

colnames(finance_data) <- c(tolower(varlist$dataname), 'year')
```

Creating Relevent Vars

Endowment Return

```
finance_data <- finance_data %>%
  mutate(endowment_ret = endowment_value_ey / endowment_value_by - 1)
```

Data Analysis

Overview

Dataset Size, Shape, Institution Coverage, Year Coverage

```
ncol(finance_data)
```

```
## [1] 82
```

```
nrow(finance_data)
```

```
## [1] 28300
```

```
finance_data <- finance_data %>%
  mutate(year = as.numeric(year))

min(finance_data$year)
```

```
## [1] 2004
```

```
max(finance_data$year)
```

```
## [1] 2017
```

```
length(unique(finance_data$unitid))
```

```
## [1] 2386
```

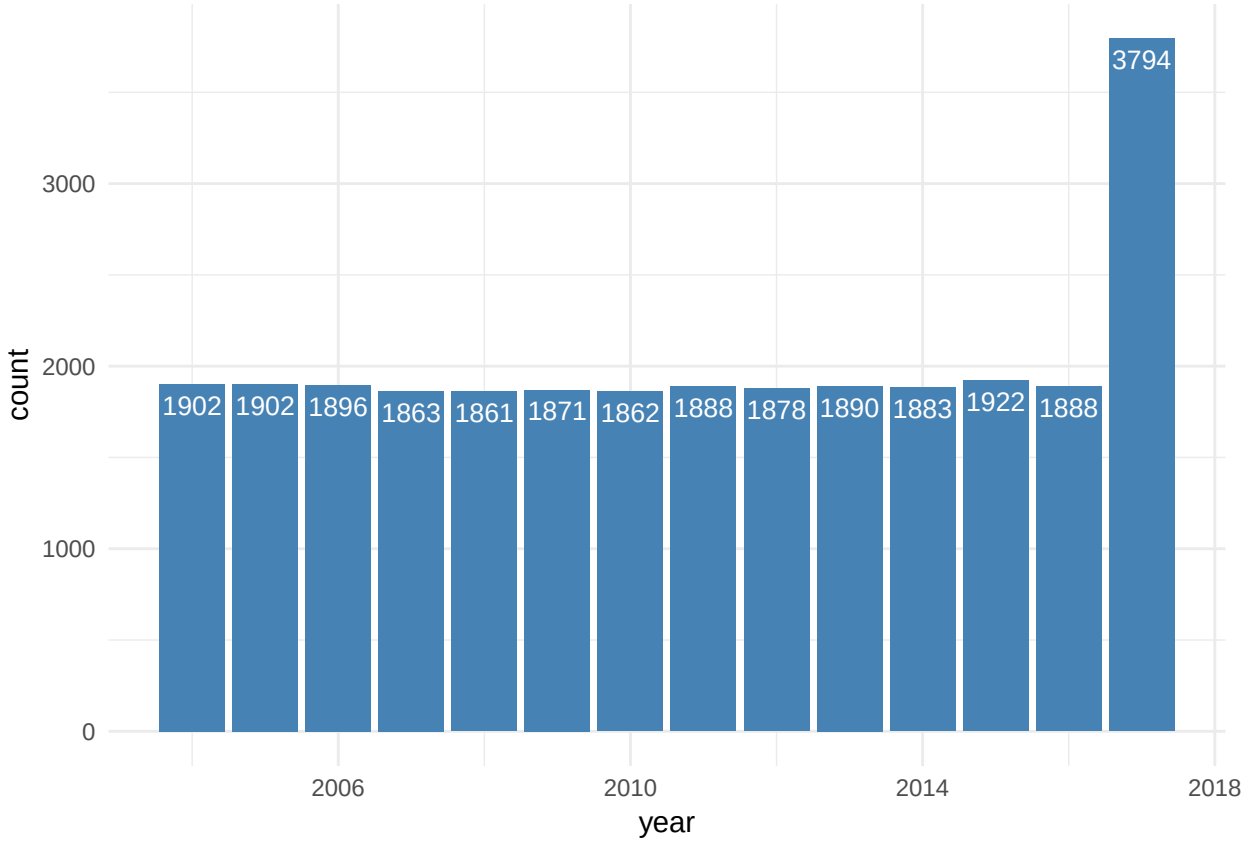
Observation count: Per institution, Per Year

```
inst <- finance_data %>%
  group_by(unitid) %>%
  summarise(count = n())
library(knitr)
kable(summary(inst))
```

unitid	count
Min. :100690	Min. : 1.00
1st Qu.:160970	1st Qu.: 9.00
Median :200540	Median :15.00
Mean :242149	Mean :11.86
3rd Qu.:243790	3rd Qu.:15.00
Max. :491057	Max. :15.00

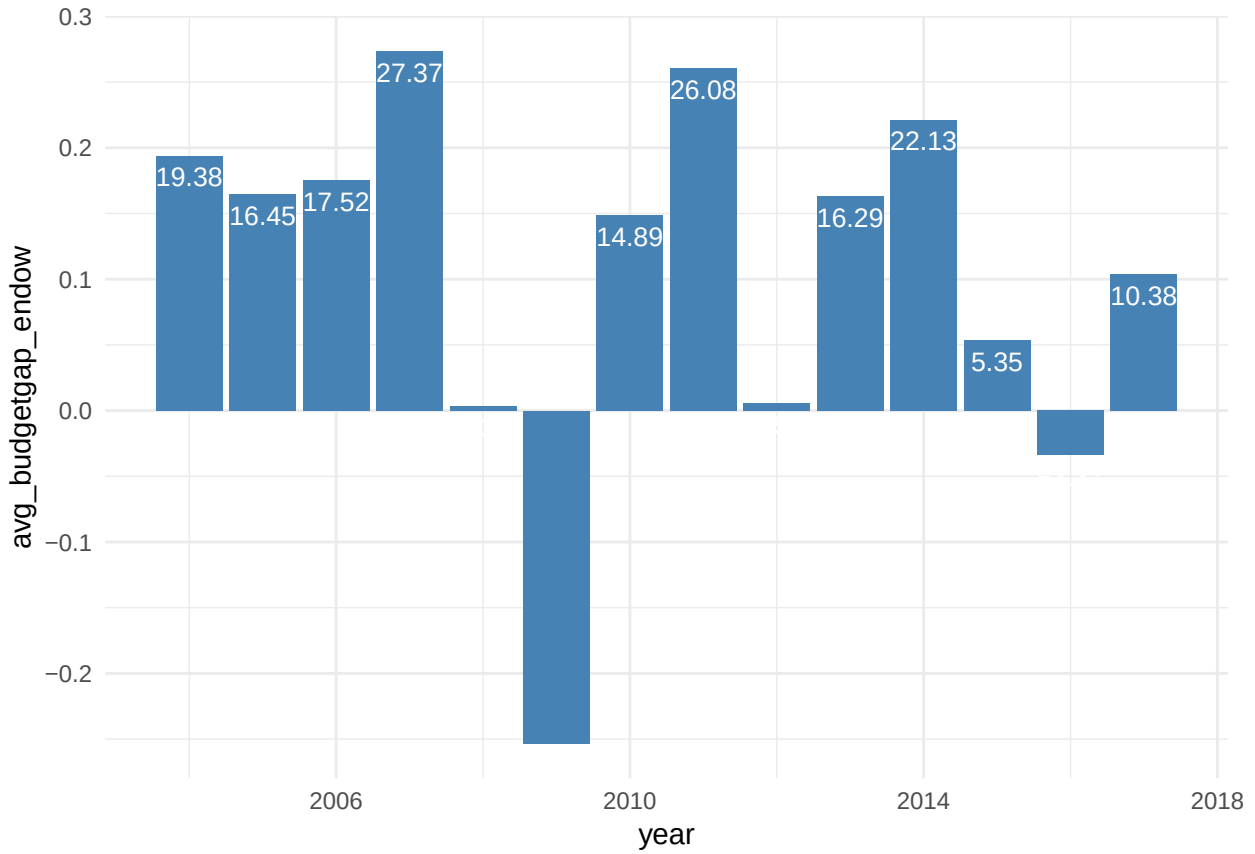
```
year <- finance_data %>%
  group_by(year) %>%
  summarise(count = n())

ggplot(data=year, aes(x=year, y=count)) +
  geom_bar(stat="identity", fill="steelblue")+
```

```
  geom_text(aes(label=count), vjust=1.6, color="white", size=3.5)+
  theme_minimal()
```

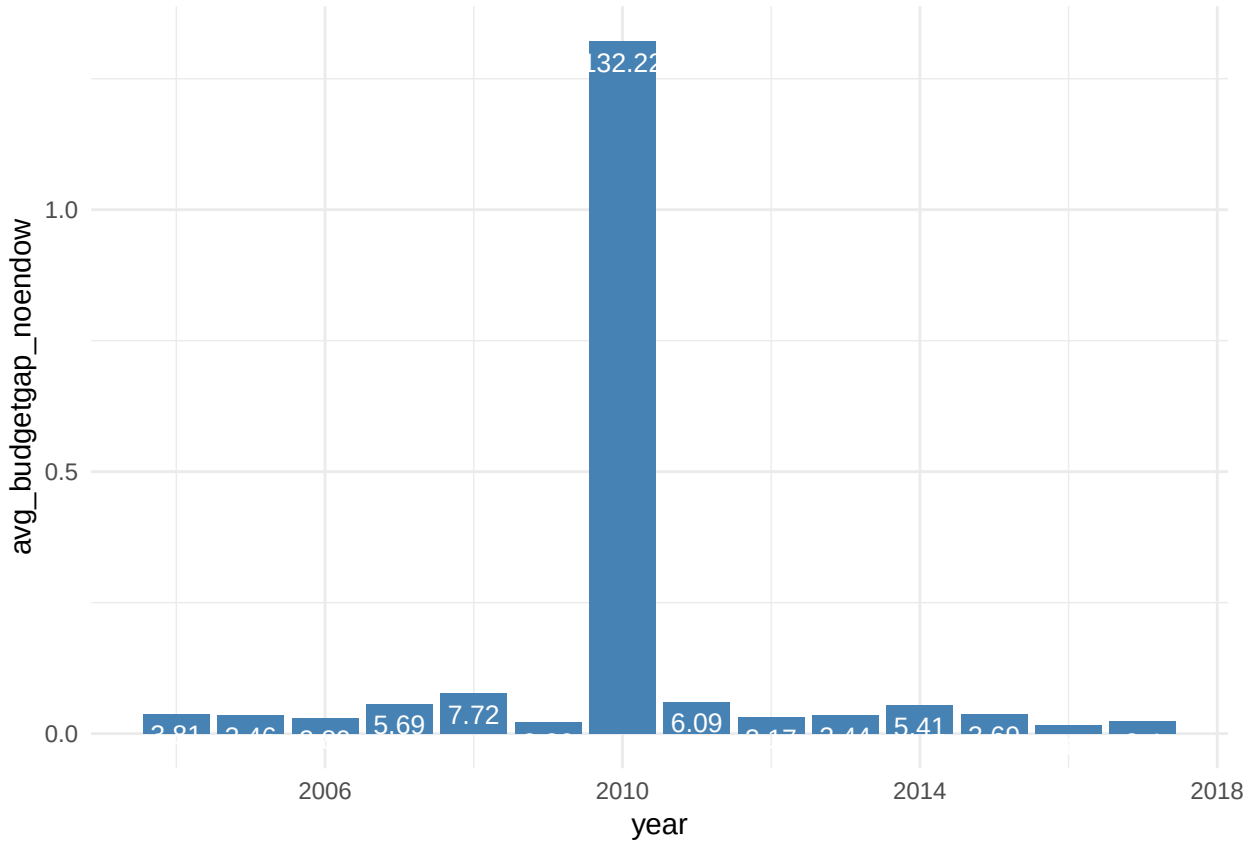


Plotting Budget Gap

```
finance_data <- finance_data %>%
  mutate(endowment_ret = ifelse(endowment_value_by == 0, 0, endowment_ret)) %>%
  filter(!is.na(endowment_ret)) %>%
  mutate(budgetgap_noendow = ((total_revenue_investment - (endowment_value_ey - endowment_value_by)) -
  mutate(budgetgap_endow = (total_revenue_investment - total_expenses) / total_expenses)

budget <- finance_data %>%
  select(total_expenses, budgetgap_noendow, budgetgap_endow, year) %>%
  filter(total_expenses > 0) %>%
  na.omit() %>%
  group_by(year) %>%
  summarize(avg_budgetgap_noendow = mean(budgetgap_noendow), avg_budgetgap_endow = mean(budgetgap_endow

ggplot(data=budget, aes(x=year, y=avg_budgetgap_endow)) +
  geom_bar(stat="identity", fill="steelblue")+
  geom_text(aes(label=round(avg_budgetgap_endow * 100 ,2)), vjust=1.6, color="white", size=3.5)+
  theme_minimal()
```

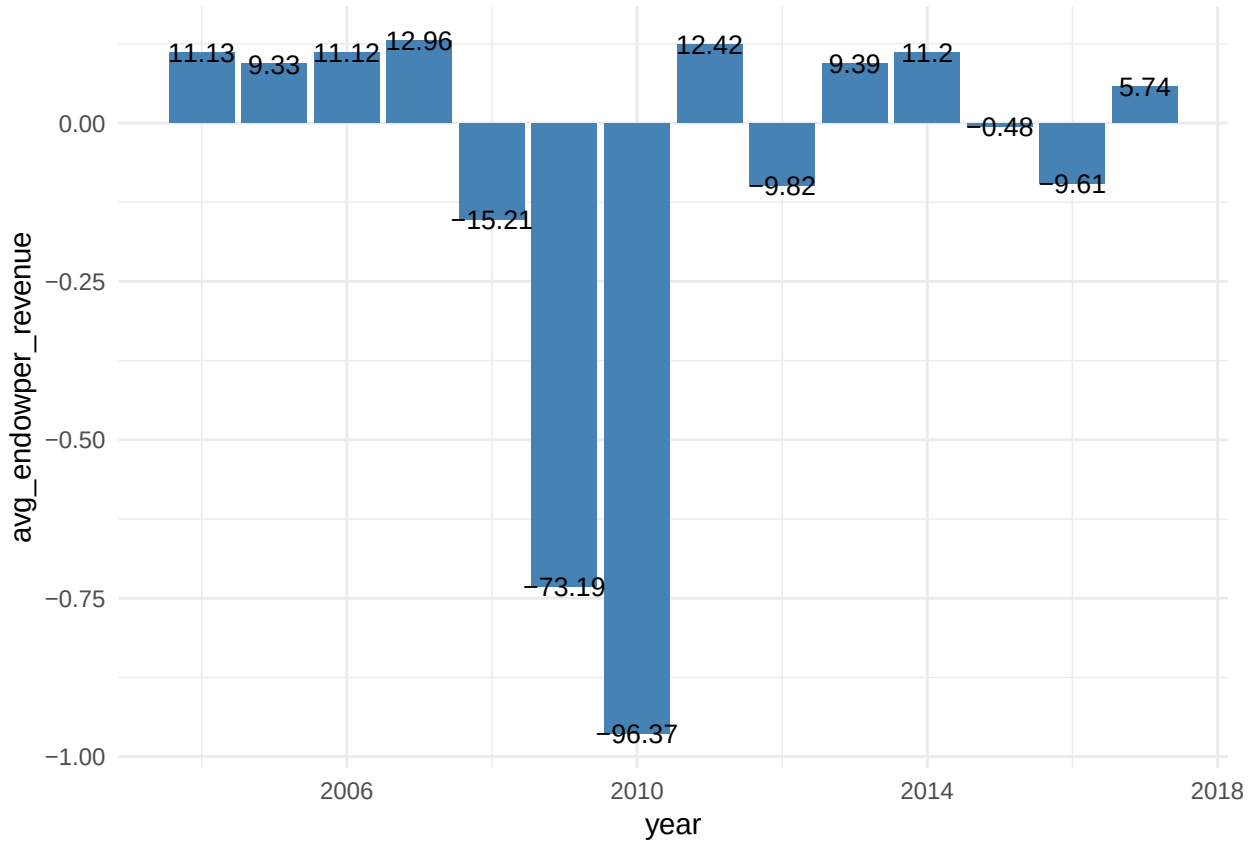



```
ggplot(data=budget, aes(x=year, y=avg_budgetgap_noendow)) +
  geom_bar(stat="identity", fill="steelblue")+
  geom_text(aes(label=round(avg_budgetgap_noendow * 100 ,2)), vjust=1.6, color="white", size=3.5)+
  theme_minimal()
```



Endowment Revenue as a % of Revenue over time

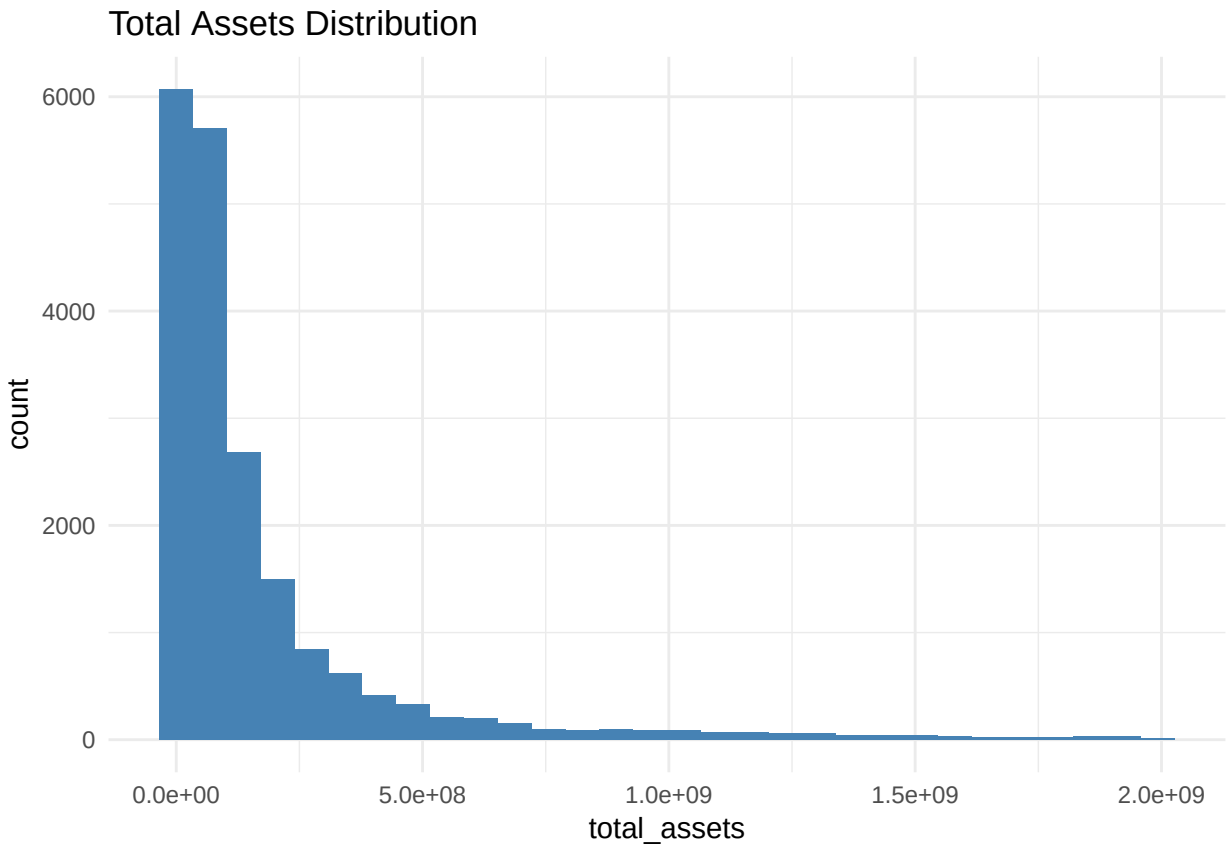
```
finance_data <- finance_data %>%
  mutate(endow_per_revenue = (endowment_value_ey - endowment_value_by) / total_revenue_investment)

endowper <- finance_data %>%
  filter(total_revenue_investment > 0) %>%
  group_by(year) %>%
  summarize(avg_endowper_revenue = mean(endow_per_revenue))

ggplot(data=endowper, aes(x=year, y=avg_endowper_revenue)) +
  geom_bar(stat="identity", fill="steelblue")+
  geom_text(aes(label=round(avg_endowper_revenue * 100 ,2)), color="black", size=3.5)+
  theme_minimal()
```



Descriptive Stats of relevant vars

```
finance_desc <- finance_data %>%
  select(unitid, total_assets, total_liabilities, total_net_assets, total_revenue_investment, total_exp

ggplot(data=filter(finance_desc, total_assets/1000000 < 2000), aes(total_assets)) +
  geom_histogram(fill = 'steelblue') + theme_minimal() + labs(title = "Total Assets Distribution")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

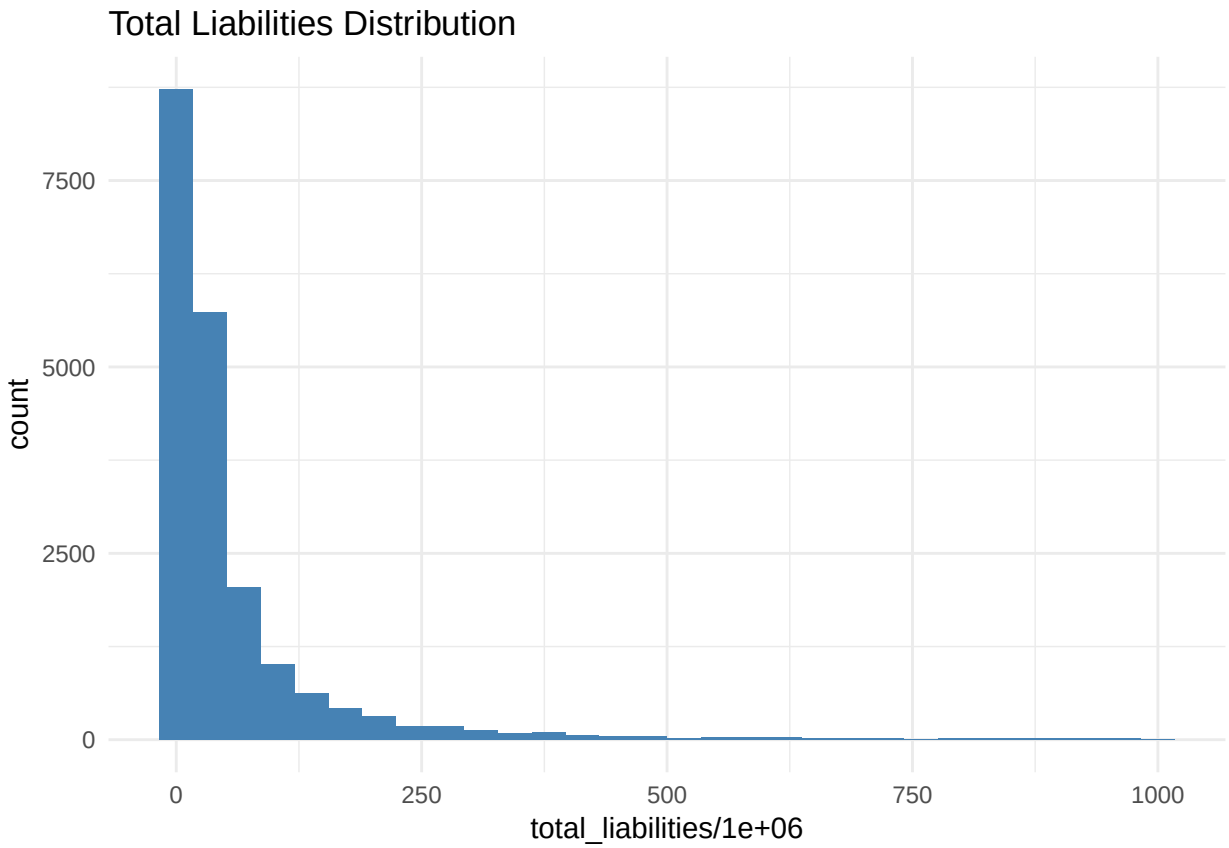



```
summary(finance_desc$total_assets / 1000000)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
##     0.00    27.69    79.57   447.43   209.31 76904.35      629
```

```
ggplot(data=filter(finance_desc, total_liabilities / 1000000 < 1000), aes(`total_liabilities` / 1000000)
  geom_histogram(fill = 'steelblue') + theme_minimal() + labs(title = "Total Liabilities Distribution")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

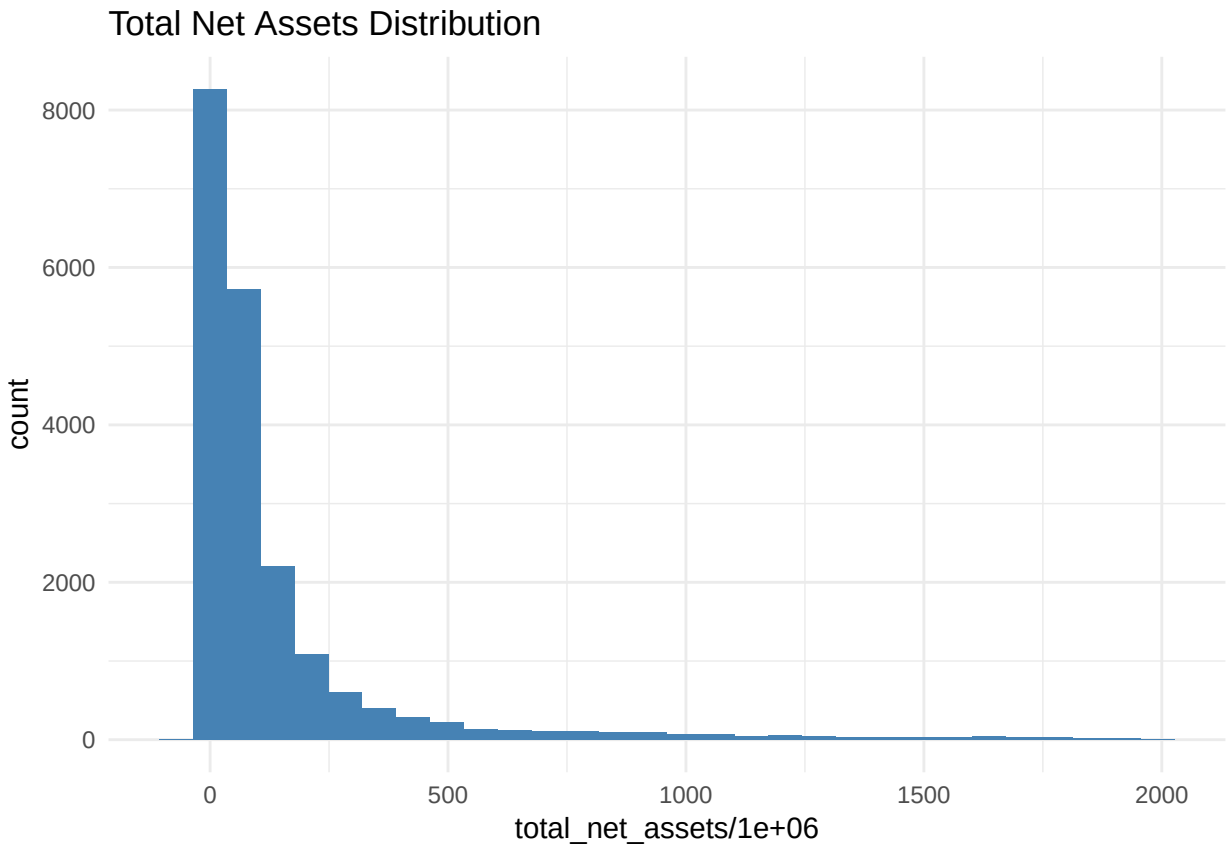



```
summary(round(finance_desc$total_liabilities))
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
## 0.000e+00 6.396e+06 2.358e+07 1.317e+08 6.285e+07 3.992e+10       629
```

```
ggplot(data=filter(finance_desc, total_net_assets/1000000 < 2000), aes(total_net_assets / 1000000)) +
  geom_histogram(fill = 'steelblue') + theme_minimal() + labs(title = "Total Net Assets Distribution")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
summary(round(finance_desc$total_net_assets))
```

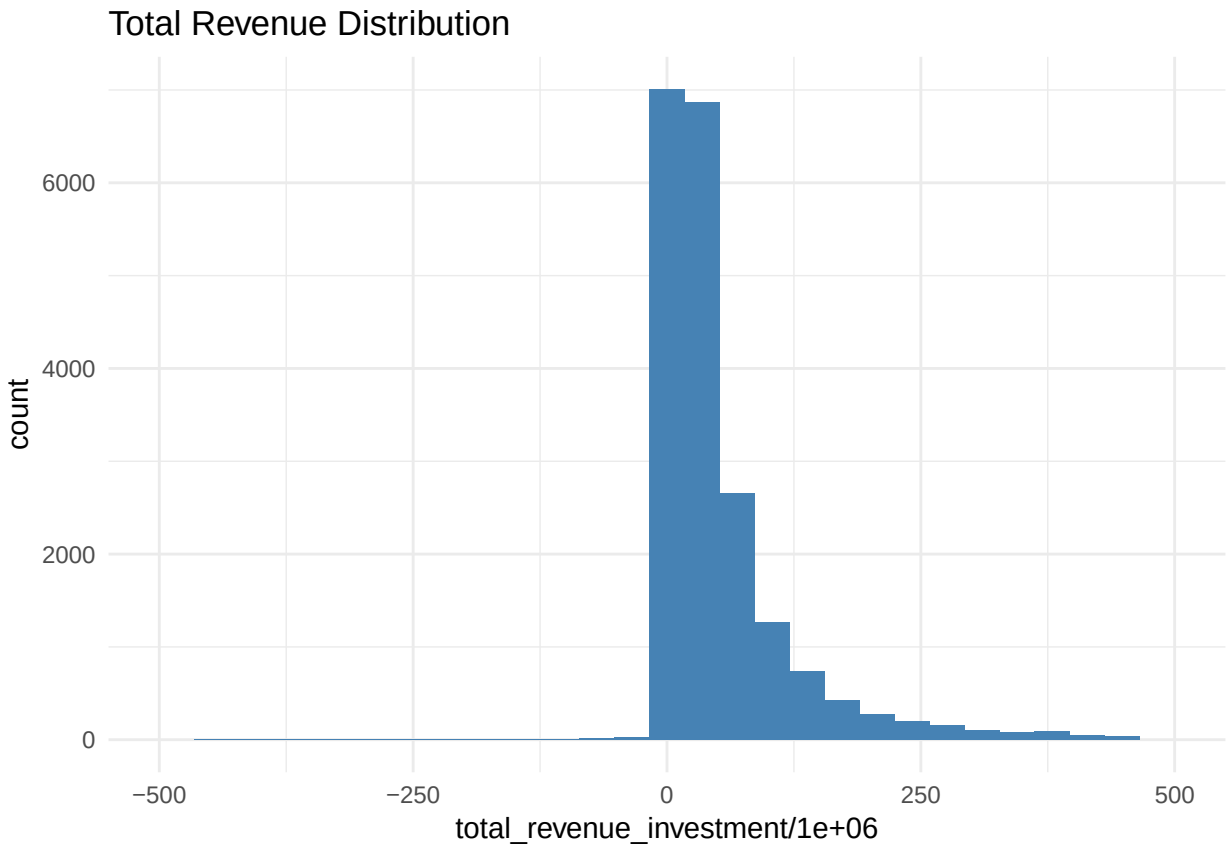
```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -6.594e+07  1.703e+07  5.147e+07  3.157e+08  1.420e+08  4.540e+10
##       NA's
##        629
```

```
ggplot(data=filter(finance_desc, total_revenue_investment/1000000 < 1000 | total_revenue_investment/1000
  geom_histogram(fill = 'steelblue') + theme_minimal() + labs(title = "Total Revenue Distribution") + xl
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 811 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```

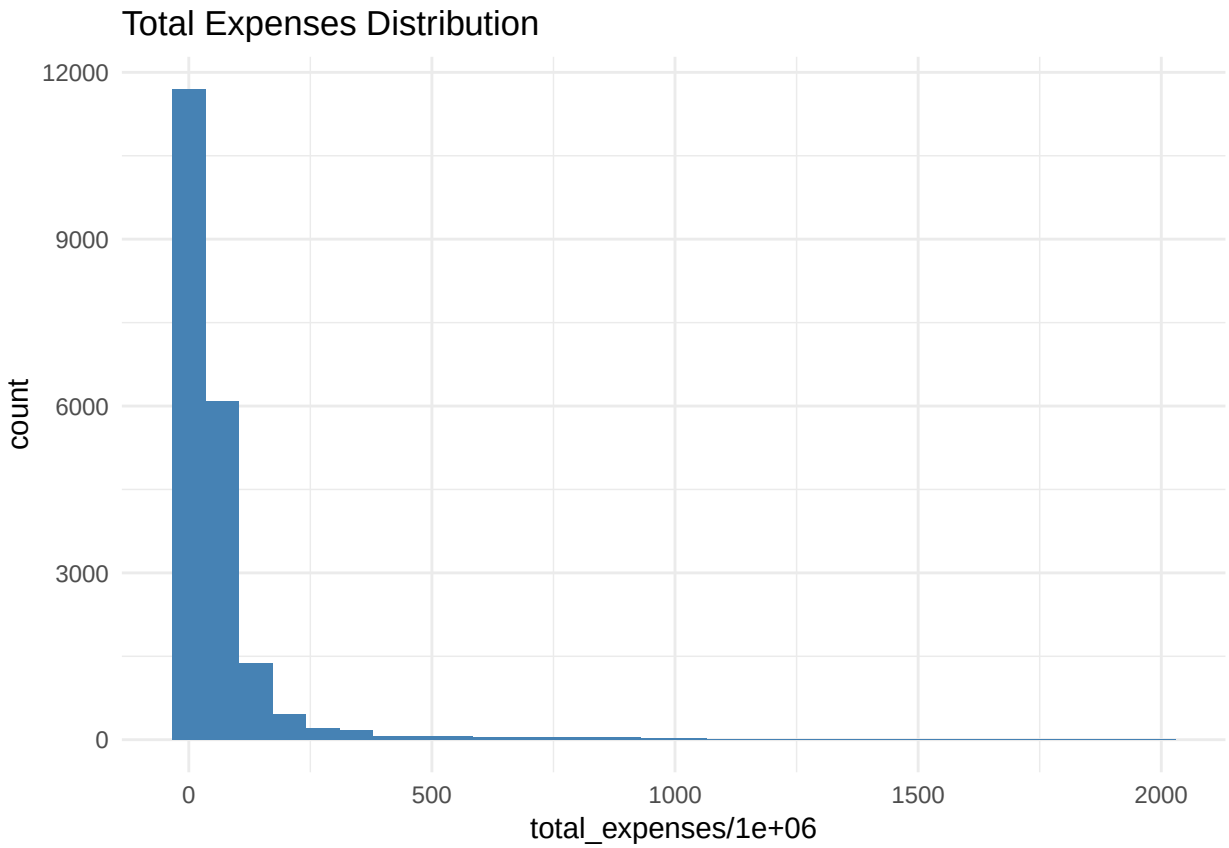



```
summary(round(finance_desc$total_net_assets))
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -6.594e+07  1.703e+07  5.147e+07  3.157e+08  1.420e+08  4.540e+10
##       NA's
##        629
```

```
ggplot(data=filter(finance_desc, total_expenses/1000000 < 2000), aes(total_expenses/ 1000000)) +
  geom_histogram(fill = 'steelblue') + theme_minimal() + labs(title = "Total Expenses Distribution")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

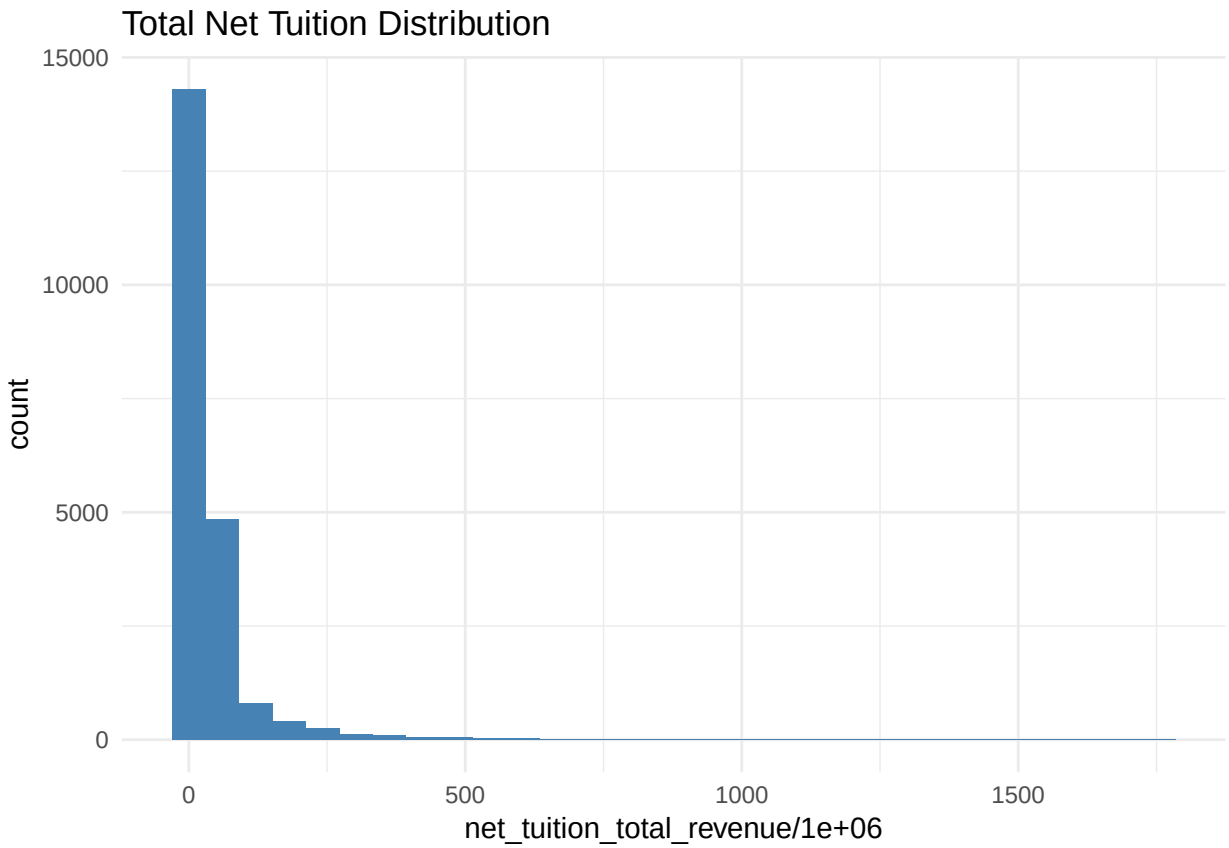



```
summary(round(finance_desc$total_expenses))
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
## 0.000e+00 1.112e+07 2.829e+07 1.101e+08 6.459e+07 8.897e+09       257
```

```
ggplot(data=filter(finance_desc, net_tuition_total_revenue/1000000 < 2000), aes(net_tuition_total_revenu
  geom_histogram(fill = 'steelblue') + theme_minimal() + labs(title = "Total Net Tuition Distribution")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

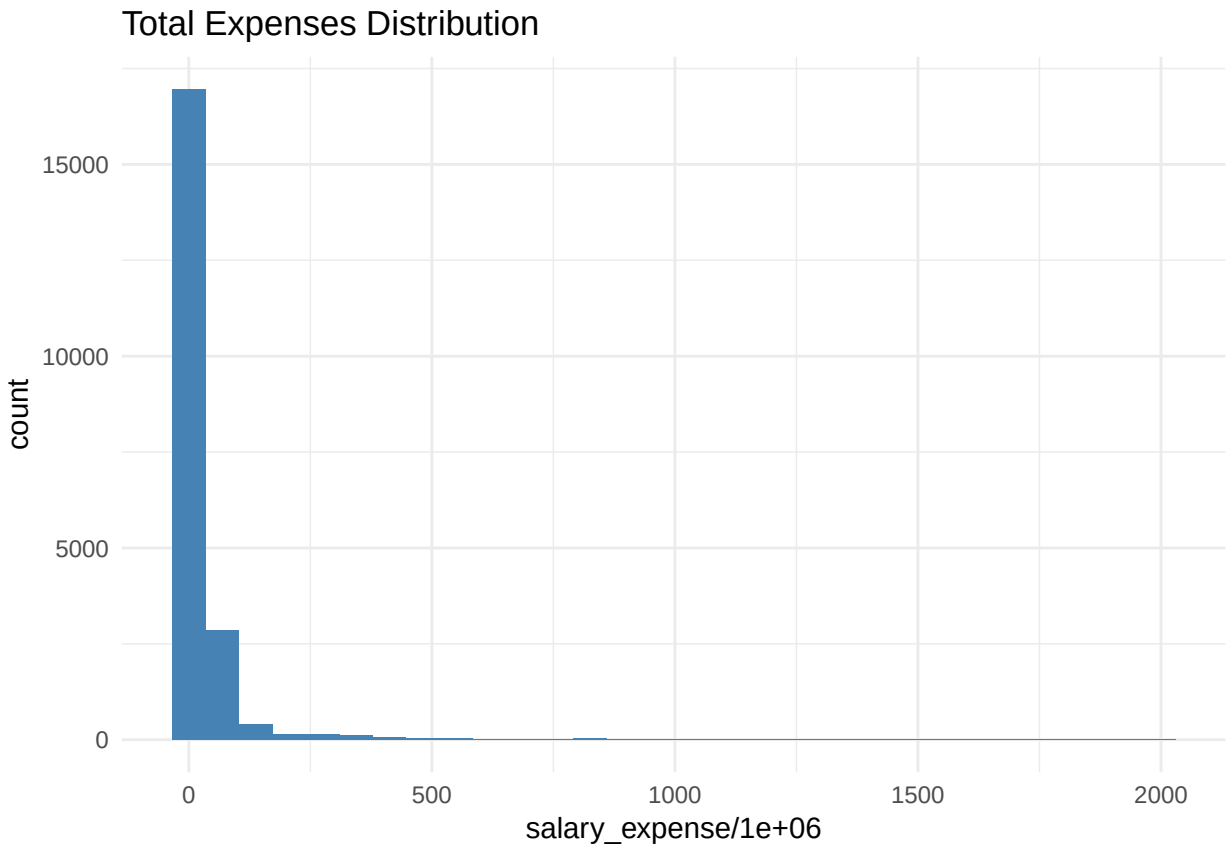



```
summary(round(finance_desc$net_tuition_total_revenue))
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## 0.000e+00 4.776e+06 1.594e+07 4.080e+07 3.895e+07 1.754e+09
```

```
ggplot(data=filter(finance_desc, salary_expense/1000000 < 2000), aes(salary_expense/ 1000000)) +
  geom_histogram(fill = 'steelblue') + theme_minimal() + labs(title = "Total Expenses Distribution")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
summary(round(finance_desc$salary_expense))
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## 0.000e+00 4.622e+06 1.192e+07 4.798e+07 2.753e+07 3.875e+09
```

```
bm_ret <- readxl::read_xlsx("6040Returns.xlsx")

bm_ret <- bm_ret %>%
  filter(year > 2004) %>%
  select(year, Blend)

finance_data <- finance_data %>%
  left_join(bm_ret, by = "year")

finance_desc <- finance_desc %>%
  left_join(bm_ret, by = "year")

finance_data <- finance_data %>%
  mutate(risk = ifelse(endowment_ret>Blend & Blend > 0, 1, ifelse(endowment_ret < Blend & Blend < 0, 1,0

length(finance_data$risk[finance_data$risk == 1]) / length(finance_data$risk)
```

```
## [1] 0.4076231
```

```
table(finance_data$risk)
```

```
##
##     0     1
## 12511  7179
```

budget gap vs risk/no risk

```r
finance_data$budgetgap_noendow[is.infinite(finance_data$budgetgap_noendow)] <- NA
finance_data$budgetgap_noendow[is.nan(finance_data$budgetgap_noendow)] <- NA

budget <- finance_data %>%
  filter(!is.na(risk)) %>%
  filter(!is.na(budgetgap_noendow)) %>%
  group_by(risk) %>%
  summarise(avg_budgetgap = mean(budgetgap_noendow))


ggplot(data=budget, aes(x=risk, y=avg_budgetgap)) +
  geom_bar(stat="identity", fill="steelblue")+
  geom_text(aes(label=round(avg_budgetgap * 100 ,2)), vjust=1.6, color="white", size=3.5)+
  theme_minimal()
```