



*Financial Review : Creating
some draft visualizations from
the data of 500 Companies, by
addressing data discrepancies
, to show their finances and
growth.*

NAME- VIVEK KUMAR SINGH

ENR.NO-A90555919025

BSC STATS VI

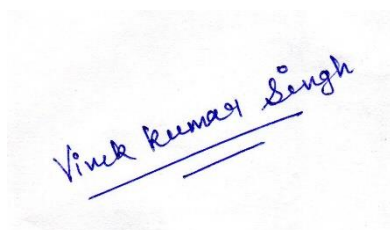
NTCC PROJECT 2022

DECLARATION

*I Vivek Kumar Singh student of B.Sc Statistics, Sem-6, hereby declare that **this seminar/dissertation/summer internship/major project/dissertation** entitled "**Financial Evaluation**" which I submit to the Department of Statistics, Amity Institute of Applied Sciences, Amity University, West Bengal, Kolkata, is partially fulfilling a requirement for the award of a Bachelor of Science in Statistics (**Honours**), which was not previously the basis for conferring a degree, diploma or similar title. other or recognition*

DATE: 26/05/2022

STUDENT SIGNATURE:

A handwritten signature in blue ink that reads "Vivek Kumar Singh". The signature is written in a cursive style and is underlined with two parallel lines.

ABSTRACT

In the world of Data Science, about 70 percent of the time goes into our data preparation, and only 30 percent into analysis, in statistics we have learned a lot about different kinds of analysis to apply depending upon the kind of data, but that is only 30% of all work, and how mind-blowing is that data preparation takes up so much of time.

Yes, of course, we've been learning and that has contributed to the fact that it has taken quite a while to go through our preparation but that is the reality of life in the world of data science.

Data preparation makes sure that our data is robust. We wanted to have non-normalized, we want to make sure everything is correct before we proceed with the analysis and that also makes the analysis very smooth and effortless.

INTRODUCTION

In this project we'll visualize the dataset of 500 Companies and based on our analysis, we'll draw useful insights but before that our dataset contains numerous discrepancies which we need to fix before proceeding with our analysis. So, we'll first write codes in R for fixing numerous discrepancies and we do this because we want our dataset to be robust. After all, it makes analysis smooth and effortless and we'll also see how to streamline the process of data preparation so that we spend less time there and will make fewer errors and we have less room for error because we know what we are doing.

We'll also want to investigate the potential threats that we're faced with, when we prepare data and in this project, we're going to see a lot about locating missing data and how to find it, how to replace missing out what to do about it.

We also have a few methods of placing missing data. one of the more advanced methods will be the main method i.e the median imputation method is a very powerful technique that is used in our programming for different types of analytics.

Dealing With Missing Data

Missing data is a very common theme in data science or analytics and it happens because the dataset is sometimes not full or there might have been some errors, while it was being supplied or somebody forgot to put something in or just basically that data was not collected in the first place.

So, dealing with missing data is important. We need to know what options we have and what approaches we can take to fix the data or maybe not fix the data and what we do.

Every different case today will overview.

1. Predict With 100% Accuracy: *So first option is the best option when you can predict the missing data with 100 percent accuracy.*

2. Leave Record as it is: *The next option is to leave the record as it is. So, don't change anything about it just leave that cell empty and when we use this approach would work if, for instance, that specific field is not important for our analysis and if the field is not important for analysis then doesn't matter that it's empty.*

The other option is that if the algorithm that you're going to be applying to your dataset and this is a bit more advanced if the algorithm you're going to be applying already incorporates some methodology is to account for missing data and to correct that inside the algorithm. And in that case, also leaving the record is OK because the competition on method will take care of it

3. Remove Record Entirely: *The Third option is to remove the record entirely so sometimes we're missing some critical data.*

And in that case when we cannot restore the data the only option that you might have left is to completely remove that whole row from your analysis to keep your analysis.

So, to keep your algorithms working or keep your approach working. You have to remove this one record.

And this has a drawback.

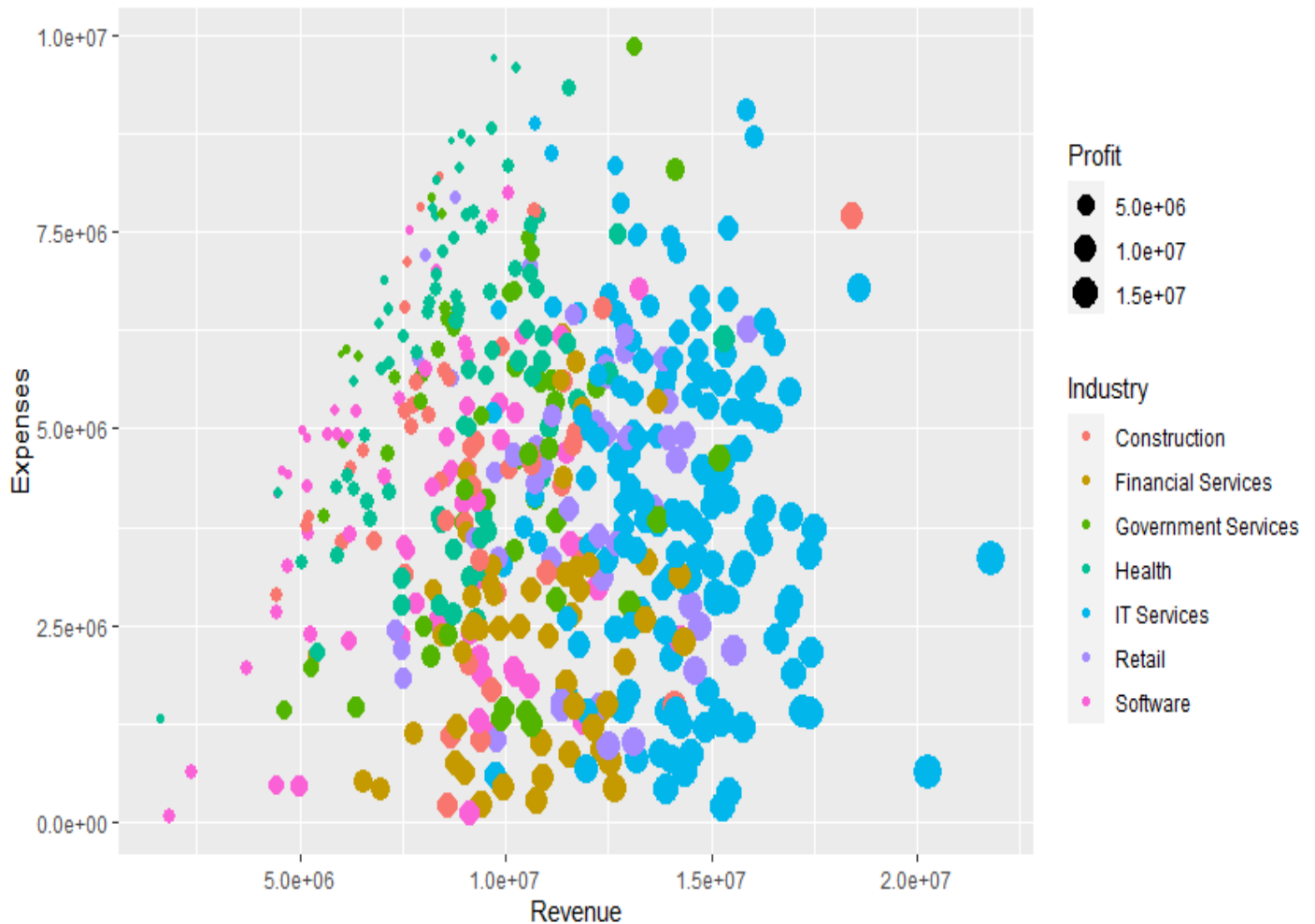
The drawback here is that your analysis becomes less significant or has a smaller sample and that can have certain implications.

So, if you're fine with sacrificing a few rows here and there and it's not critical for you to have information and every single record then this could be another approach that you might apply.

4. Replace with mean or median: *Next one is a replacement with the mean or median.*

This is a very popular approach and we're going to also see an example of this in our dataset in our project this is going to be one of the main approaches we're going to be relying on. And so basically here we can say that if for instance you have lots and lots of rows and they all have a certain value for revenue and that one row is missing that value then you can calculate the mean or the median and put it in there and then proceed with your analysis. So, Mean is an ok option but usually, the Median is preferred because the median is less affected by outliers

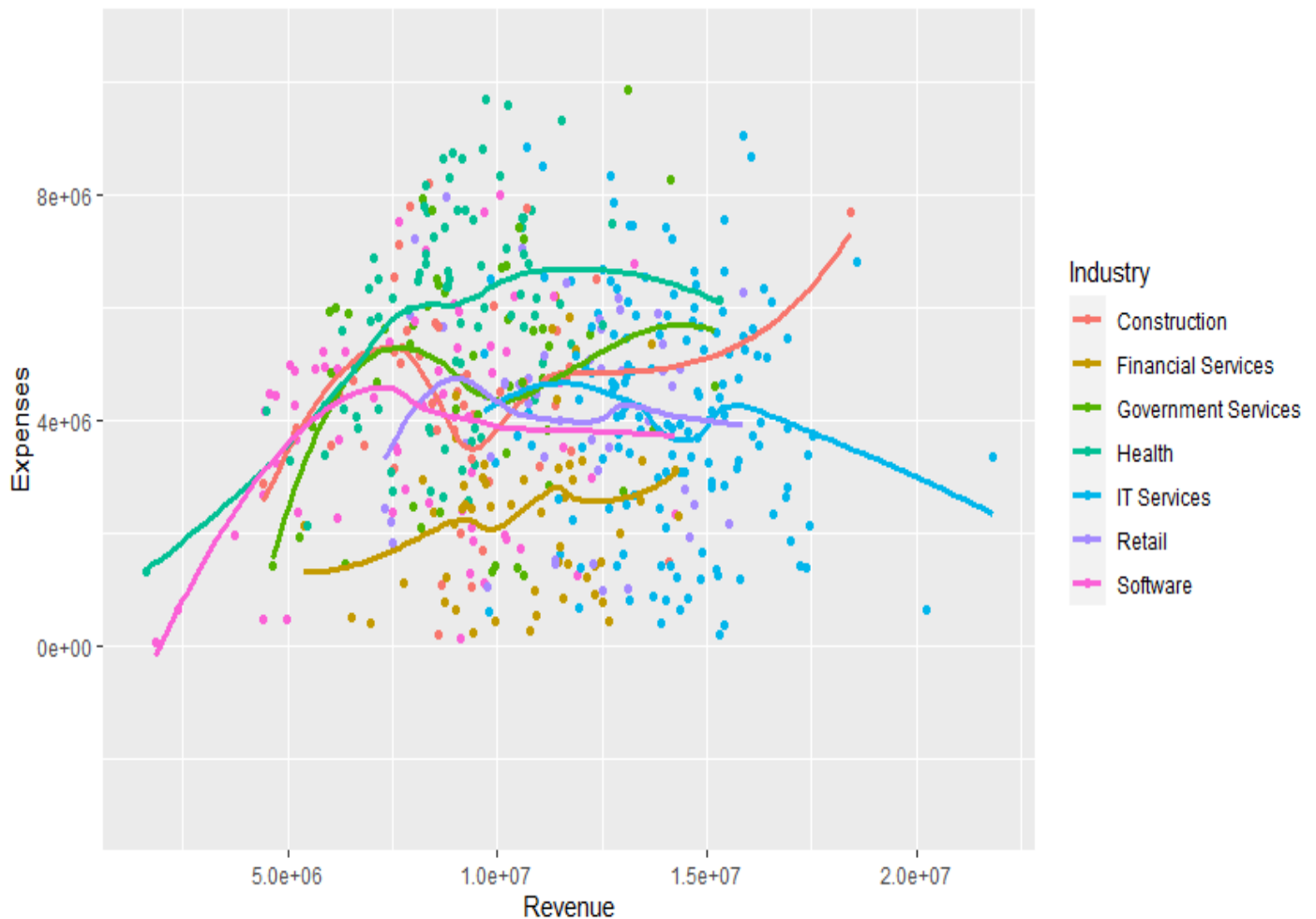
IMPORTANT VISUALIZATIONS



*From this graph, we can tell that the majority of **IT Services** are making a good amount of revenue even at low expenses.*

*Meanwhile, **Government Services** are the ones with low revenue even at high expenses.*

*While **Software Companies** don't have a consistent record.*



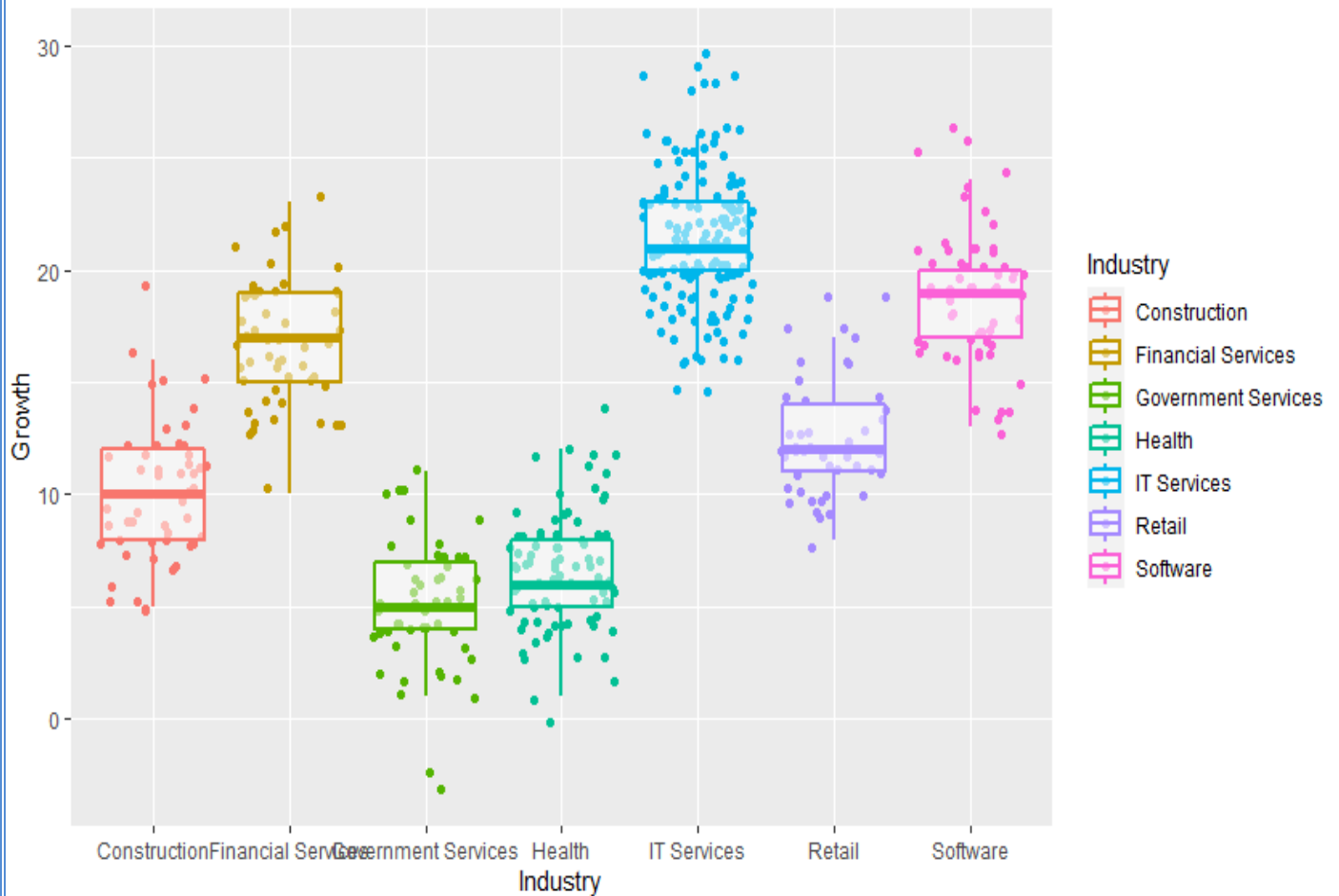
And in this plot, we can see some interesting trends. So, for instance, we can see construction companies it's going like it's got this big dip, quite a few companies constructions and government services But as construction companies and government services companies kind of start nearing the 10 million dollar mark in revenue their expenses start increasing

which eats into their revenues and it starts decreasing. Interesting enough.

So obviously this is the best part to be where your revenues growing expenses are decreasing. But then as revenue keeps growing further for some companies expenses start increasing for other companies on the other hand expenses start decreasing.

And In the case of financial services, we see quite a consistent trend for the financial services companies so that consistency speaks to probably that they know what they're doing generally.

So they're all kind of on the same trajectory.



*In this boxplot, we can see **the construction industry** is sitting at around 10 percent in growth as we saw.*

*And then we've got some other companies some industries **like I.T. services** is up there at the highest.*

it's government services that are the lowest.

TAKEAWAYS

- *IT Services generate a good amount of revenue even at low expenses.*
- *Government Services generate low revenue as compared to other sectors even at high expenses*
- *Construction companies and Government services as they kind of start nearing the 10 million dollar mark in revenue their expenses start increasing which eats into their revenue and it starts decreasing.*
- *Financial Services are consistent in terms of revenue.*

- *IT Services has the highest growth as compared to other sectors*
- *Government Services are the ones with the lowest growth.*

CODES FOR DATA PREPARATION AND VISUALIZATION

```
getwd()
setwd("C:/Users/singh/Downloads")
getwd()
#Basic : fin <-
read.csv("Future500.csv")
fin <-
read.csv("Future500.csv", na.strings
= c(""))
fin
head(fin)
tail(fin)
str(fin)
#factors
fin$ID <- factor(fin$ID)
fin$Inception <-
factor(fin$Inception)
summary(fin)
#factor variable trap
#a <- c("12", "13", "14", "12", "12")
```

```
#a  
#typeof(a)  
#b<-as.numeric(a)  
#b  
#typeof(b)  
#Converting into numeric for  
factors  
#z<-  
factor(c("12", "13", "14", "12", "12"))  
#z  
#k <- as.numeric(z)  
#k  
#z  
#p<-as.numeric(as.character(z))  
#p  
head(fin)  
str(fin)  
head(fin)  
fin$Expenses <- gsub("Dollars", "", fin$Expenses)  
head(fin)  
fin$Expenses <-  
gsub(",", "", fin$Expenses)
```

```
str(fin)
fin$Revenue <-
gsub("\\$", "", fin$Revenue)
fin$Revenue <-
gsub(",", "", fin$Revenue)
fin$Growth <-
gsub("%", "", fin$Growth)
head(fin)
str(fin)
fin$Expenses <-
as.numeric(fin$Expenses)
fin$Revenue <-
as.numeric(fin$Revenue)
fin$Growth <-
as.numeric(fin$Growth)
str(fin)
summary(fin)
#Locating Missing Values
head(fin, 24)
complete.cases(fin)
fin[!complete.cases(fin),]
is.na(fin$Expenses)
fin[is.na(fin$Expenses),]
```

#.....Removing Records with missing data

```
fin_backup <- fin
fin_backup
fin[!complete.cases(fin),]
fin[is.na(fin$Industry),]
fin <- fin[!is.na(fin$Industry),]
fin
fin[!complete.cases(fin),]
```

#Resetting the data frame index

```
fin
rownames(fin) <- 1:nrow(fin)
fin
rownames(fin)<- NULL
tail(fin)
```

#Replacing Missing Data: Factual Analysis

```
fin[is.na(fin$State),]
fin[is.na(fin$State) &
fin$City=="New York", "State"] <-
"NY"
fin[c(11,377),]
```

```

fin[is.na(fin$State) &
fin$City=="San Francisco","State"]
<- "SF"
fin[is.na(fin$State),]
fin[c(82,265),]
fin[!complete.cases(fin),]
#Replacing Missing Data: Median
Manipulation
fin[!complete.cases(fin),]
#median(fin[, "Employees"])
#median(fin[, "Employees"], na.rm =
TRUE)
#median(fin[fin$Industry=="Retail",
"Employees"], na.rm=TRUE)
#mean(fin[fin$Industry=="Retail", "E
mployees"], na.rm=TRUE)
med_emp_retail <-
median(fin[fin$Industry=="Retail", "
Employees"], na.rm = TRUE)
med_emp_retail
fin[is.na(fin$Employees) &
fin$Industry=="Retail", "Employees"]
<- med_emp_retail

```

```
fin[3,]  
med_emp_fs <-  
median(fin[fin$Industry=="Financial  
Services", "Employees"], na.rm =  
TRUE)  
med_emp_fs  
fin[is.na(fin$Employees) &  
fin$Industry=="Financial  
Services", "Employees"] <-  
med_emp_fs  
fin[!complete.cases(fin),]
```

#Replacing Missing Data: Median Manipulation 2

```
med_gr_cst <-  
median(fin[fin$Industry=="Construct  
ion", "Growth"], na.rm = TRUE)  
fin[is.na(fin$Growth) &  
fin$Industry=="Construction", "Growth"] <- med_gr_cst  
fin[!complete.cases(fin),]
```

#Replacing Missing Data: Median Manipulation 3

```

med_rv_cst<-
median(fin[fin$Industry=="Construct
ion", "Revenue"],na.rm = TRUE)
med_rv_cst
fin[is.na(fin$Revenue) &
fin$Industry=="Construction", "Reven
ue"] <- med_rv_cst
fin[!complete.cases(fin),]
med_ex_cst<-
median(fin[fin$Industry=="Construct
ion", "Expenses"],na.rm = TRUE)
med_ex_cst
fin[is.na(fin$Expenses) &
fin$Industry=="Construction", "Expen
ses"] <- med_ex_cst
med_ex_cst
fin[!complete.cases(fin),]
#med_ex_cst<-
median(fin[fin$Industry=="Construct
ion"&
is.na(fin$Profit), "Expenses"],na.rm
= TRUE)
#med_ex_cst

```

```

#fin[is.na(fin$Expenses) &
fin$Industry=="Construction" &
is.na(fin$Profit), "Expenses"] <-
med_ex_cst
#med_ex_it<-
median(fin[fin$Industry=="IT
Services", "Expenses"], na.rm = TRUE)
#fin[is.na(fin$Expenses) &
fin$Industry=="IT
Services", "Expenses"] <- med_ex_it
fin[!complete.cases(fin),]
#Replacing Missing data : deriving
Values
fin[is.na(fin$Profit), "Profit"] <-
fin[is.na(fin$Profit), "Revenue"]-
fin[is.na(fin$Profit), "Expenses"]
fin[!complete.cases(fin),]
fin[is.na(fin$Expenses), "Expenses"]
<-
fin[is.na(fin$Expenses), "Revenue"]-
fin[is.na(fin$Expenses), "Profit"]
fin[!complete.cases(fin),]
fin

```


#....Visualizations

#A scatterplot classified by industry showing revenue, expenses, profit

#A scatterplot that includes industry trends for the expenses~revenue relationship

#BoxPlots showing growth by industry

library(ggplot2)

p <- ggplot(data=fin)

p

p + geom_point(aes(x=Revenue, y=Expenses, colour=Industry, size=Profit))

A scatter plot that includes industry trends for the expenses

d <- ggplot(data=fin, aes(x=Revenue, y=Expenses, colour=Industry))

d + geom_point() +

```
geom_smooth(fill=NA, size=1.2)
```

#Boxplots

```
f <- ggplot(data = fin,  
aes(x=Industry, y=Growth,  
colour=Industry))  
f + geom_boxplot(size=1)
```

#Extra

```
f + geom_jitter() +  
geom_boxplot(size=1,  
alpha=0.5, outlier.colour = NA)
```