Factors that inform SMME profitability in South Africa using the 2010 FINSCOPE dataset

Lindokuhle Tshongolo

2023-04-13

Introduction.

Research Purpose

South Africa faces a persistent unemployment problem, with around 34.5% of the adult and working population unemployed. Figures are even more concerning when we look at youth unemployment which stands at a staggering 63.9% for the young between the ages of 15-24 and around 42.1% for those who are between the ages of 25 - 35. These numbers include the people who are eligible to work and are actively looking for work. Such high numbers have potentially devastating effects on the people in society, who are affected both directly and indirectly. Issues such as crime, poverty, and lack of social cohesion can arise as a result of such high unemployment figures. Thus potential solutions to unemployment are of utmost importance to government and policy makers.

SMMEs (Small Micro and Medium Enterprises) have been pointed out as being the potential solution to the unemployment issue. SMMEs refer to enterprises with around 0 - 200 employees. Those with 0 employees are referred to as own-account, 1-10 as micro, 10-49 as small and those with 50-200 employees as Medium enterprises. The identification of SMMEs as being the potential solution can be attributed to the fact that SMMEs have low barriers of entry and are accessible to many with little capital. Furthermore, these enterprises have the potential of introducing innovation and bringing about new ideas to the economy which the bigger businesses might have missed. Moreover, SMMEs have the ability to go after the smaller opportunities in the economy which bigger companies can't go after as this might distract them from their core businesses, and capitalize on them and thus bring about growth and employment. Likewise, SMMEs have the advantage of being local and thus have the ability to go after the local opportunities and offer solutions to their immediate communities, again having a competitive advantage over the bigger businesses that are removed from the ground and operate in a much larger area. Lastly, bigger companies can delegate some of their activities to SMMEs as a form of cost-saving mechanism. This can be the same for the government, which can offer relatively smaller tenders or subcontract work to these SMME enterprises.

However, SMMEs have high failure rates, with about 75% of all SMMEs failing within the first two years of being founded. This reveals that SMMEs tend to face a lot of challenges in the markets and thus pose a serious challenge to the efforts made by the government and its policymakers in using SMMEs as a solution to unemployment. Therefore, much work needs to be done to ensure their success, and the factors that contribute to their failure or

success need to be known such that the appropriate support measures can be given to new founders and thus contribute to their growth.

In this analysis, we will be looking at the factors that drive the success or failure of SMMEs in the South African context using the 2010 Finscope SMME data to analyze these factors that contribute to the success of SMMEs. The target variable we will use as the proxy for SMME performance is business monthly profits and a slew of other variables will be used as the predictor variables.

Technical Steps

This analysis first selects all the columns that are deemed relevant using mostly the literature associated with the SMME sector in South Africa, and thus no statistical methods will be used to do the initial variable selection. This was due to a variety of reasons, one being a lack of resources both in terms of technical skills and time considerations, with having to clean a total of about 2012 variables of which many were repetitive and thus had the potential of introducing redundancy. This initial selection of variables will reduce the number of variables from the original data set from 2012 variables to about 52 variables which is still a tall order but a manageable task.

The second aspect of our analysis and the most elaborate will deal with the creation of new variables that will be useful in our analysis. Note, that the inclusion of many variables initially is to help the model deal with biases in trying to understand the factors that determine business performance. Too few variables might introduce biases into our model as the model might not be able to understand the nuances, variability, and sources of such variability in the data. Further, we will also clean the existing variables such that they reflect their appropriate data types. Again, this analysis will motivate the importance of such variables that have been included in the model and how they will aid in our analysis. Then we move on to the cleaning of the columns to ensure that they are consistent and do not have redundancy or repeated column names.

Then, this analysis will carry out outlier detection using studentized residuals to identify them initially. To figure out if such outliers are influential, this analysis makes use of cook's distances, which is a composite measure of outliers and leverage to find influential data points. To deal with the issue of missing values, Predictive Mean Matching is used to find the appropriate values to fill in the missing values. Then, we use the VIF(Variance inflation factor) which is a ratio of 1 to 1 less the coefficient of determination, with higher ratio values associated with high multicollinearity, to find highly correlated or multicollinear variables and remove them from the model. Then, log transformations are used to fix non-linearity issues in the data and lessen the severity of heteroscedasticity if it exists. Furthermore, we use the Resset test to test for the presence of model misspecification and further make use of random forests to deal partially with the issue of model misspecification by finding important interactions using the *random forest explainer* package. As a benchmark model, this analysis will make use of the logged linear regression model. This model can solve the issue of non-linearity and partially solve the issue of heteroscedasticity. Finally, the lasso regression model will be used to find the most robust variables that have the most explanatory power when it comes to SMME performance. The

lasso regression has the ability to remove any multicollinearity and unimportant variables by sending them to zero.

Data Importing and variable naming

Here we will be importing the data from the local directory to the working environment, and loading the libraries that will be working with, the work done by each library will be specified when it is invoked at different parts of this document. The name of the dataset that this analysis will make use of is the finscopeSMME2010 data collected across South Africa, and the details of the dataset can be provided on request.

```
library(readxl)



rm()

#FinScopeSMME2010 <- read_excel("FinScopeSMME2010.xlsx")
#View(FinScopeSMME2010)
library(stargazer)   # For tables

##
## Please cite as:

##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary
Statistics Tables.

##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer

library(ggplot2)    #for the graphs

## Warning: package 'ggplot2' was built under R version 4.2.3

library(tidyverse)   #for manipulating the data

## Warning: package 'tidyverse' was built under R version 4.2.3

## Warning: package 'tibble' was built under R version 4.2.3

## Warning: package 'tidyr' was built under R version 4.2.3

## Warning: package 'readr' was built under R version 4.2.3

## Warning: package 'purrr' was built under R version 4.2.3

## Warning: package 'dplyr' was built under R version 4.2.3

## Warning: package 'stringr' was built under R version 4.2.3

## Warning: package 'forcats' was built under R version 4.2.3

## Warning: package 'lubridate' was built under R version 4.2.3
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✓ dplyr     1.1.2     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ lubridate 1.9.2     ✓ tibble    3.2.1
## ✓ purrr     1.0.1     ✓ tidyr     1.3.0

## ── Conflicts ────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors

library(dplyr)
library(tableone)
library(janitor)

## Warning: package 'janitor' was built under R version 4.2.3

##
## Attaching package: 'janitor'
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test

library(corrplot)

## Warning: package 'corrplot' was built under R version 4.2.3

## corrplot 0.92 loaded

library(car)

library(lmtest)

library(readr)
library(readxl)
library(caret)

library(car)
library(lmtest)
library(randomForestExplainer)

library(randomForest)

library(coefplot)
```

Variable Selection

To select the relevant variables from the bigger Excel file we make use of the *select* function, then use the *colnames* function to rename the columns to much more appropriate

column names which reflect the information they contain. Note, the codes of these columns correspond to the original Excel document which matches each code to a column name.

```
finscope_data_selected <- finscope_data %>%

select(`ID`,`Q920#`,`Q921#`,`Q81#`,`Q82#`,`Q86#`,`Q87#`,`Q89#P`,`Q89#Q`,`Q89#
R`,`Q89#S`,`Q89#T`,`Q89#U`,`Q89#V`,`Q257#`,`Q117#A`,`Q117#B`,`Q117#C`,`Q117#D
`,`Q117#E`,`Q121#`,`Q124#`,`Q133#AD`,`Q133#AE`,`Q133#AF`,`Q133#AI`,`Q133#AG`,
`Q153#`,`Q152#A`,`Q690`,`Q692`,`Q41#`,`Q228#`,`Q229#`,`Q230#`,`Q232#`,`Q233#`
,`Q183#`,`Q605#`,`Q624#`,`Q175#`,`Q171#`,`Q170#LN`,`Q123#`,`Q120#`,`Q128#`,`Q
184#B`,`Q231#`,`Q189#`,`Q191#`,`Q637#`,`Q180#`,`Q95#`)

colnames(finscope_data_selected) <-
c("ID1","Location","Province","Businesstype","wherebusoperate","ownrent","Tot
HoursWrk","vision/mis","busPln","busstr","mrkpln","Accfinrec","frmlTrnStff","
busbgt","TotNumWorkers","PrivIndv","otherSmallBus","OtherLargeBus","Gov","Oth
er","TenderSucc","RegWthCIRPO","BusContInsOffEquip","BusContSpecialisedToolso
rMchinery","PrpStrBusPremIns","CrpIns","AccDamTransIns","LargeSrcBorr","BankL
oan","BusTurnMonthly","BusNetProfmnthly","Age","Gender","Race","Mar
Stat","HighlevelEdu","IsBusOnlySrcInc?","KeepFinRec","OvrallFinAccess","Credi
tBorrowingStrd","HavingSecurityMeasures","sufferedCrimeortheft","ClaimInsforT
heft","IsBusRegistered","SubmittedTenderProp",
"BEEContrStatScoreCrd","CompFinRecs","OwnerRentorOwnPrivRes","ExptoCUST","Sup
pOutofSA","HaveIns","OffGoodCred","YearBusStart")
```

Now that we have selected all the necessary columns, and renamed them to much more appropriate names, though it might seem that the names are still codified, however, with the new column names, the reader can follow the names of these columns much more intuitively. We can now move to a new part of our analysis, the creation of new variables.

Creating variables

Then, this stage of the analysis pertains to the creation of new variables and cleaning up the existing ones so that they match their relevant data types, i.e. Categorical, numerical variables, etc. In this part of the analysis, we will also create new variables that will aid us in our analysis. For each variable that will be created in this part, an explanation of what each portion of the code is doing will be provided as much as possible whilst simultaneously trying to limit redundancy. Further, a motivation of why such a variable is necessary and how it will aid us in our analysis will be extensively provided.

The Target Variable and Motivation for the study

The first variable we will look into is the explained/dependent variable in our analysis which pertains to business performance. From the data set, there are two variables that can be good proxies for business performance, which are, *business monthly turnover* and *monthly profits*. However, these two variables are most likely to be highly correlated, thus we will choose monthly profitability as a proxy for business performance. The logic for choosing businesses that are profitable as a proxy owes to the fact that profits are net expenses as opposed to revenue which just look into money that comes into a business.

Therefore, we argue that net expense income is a much more robust indication of business performance than just pure revenue, as revenue can be drained by expenses in a business such as operating expenses, liabilities, or high labor costs, thus leaving a small portion as profits. Thus, shielding appropriate business performance.

As outlined and motivated above, this analysis uses *business monthly profits* as the explained/dependent variable which we seek to understand. All the other variables in this analysis are added as potential robust explanatory variables for profitability, thus ultimately the performance of SMMEs (Small Micro Medium Enterprises). Identification of such variables can ultimately be of use to investors or creditors as an understanding of the factors that determine profitability can aid these stakeholders in picking good businesses thus leading to a return on investment or the honoring of debts. In the case of policymakers, this framework can aid them in coming up with the appropriate support for small enterprises as they would have a firm understanding of those factors that are important in determining success. This analysis and its results can also be of value to small business owners, as a proper understanding of the factors that are significantly associated with performance can aid them in planning how to structure their businesses such that they are profitable.

The cleaning of this variable was an elaborate affair. There was some attrition by the respondents who refused to disclose how much they were making for their business in turnover on a monthly basis. Other respondents simply didn't know how much they were making to begin with while others refused to give a figure for their incomes. Thus, to deal with these missing data points, we assigned *refused* and *Don't know* to *NA* which is a statement for missing values. By assigning to missing, we are implicitly assuming that the missing values or the attrition was happening at random and there was no underlying pattern as to why some members attrition-ed, otherwise if there was a pattern then that would bring bias to our model. We then used numeric to turn the variable to its appropriate data type of numeric. The variable of Monthly profits also follows the same trajectory as monthly turnover, though the pattern of missing values might be different, i.e. no two columns in monthly turnover and monthly profit might be missing at the same time.

```
finscope_data_useful <- finscope_data_selected #This portion is to simple
make a new dataframe and copy the data from the initial one.

#Income (turnover)/ THERE is a lot of missing values, I wish whatever model I
pick to imputate the missing values doesnt't rely on mean since the mean is
unreliable source of predicting missing values..
unique(finscope_data_useful$BusTurnMonthly)
```

```
##  [1] "12500"      "220000"     "16000"      "7500"       "1500"
##  [6] "7000"       "Don't know" "5000"       "200"        "700"
## [11] "375"        "750"        "10000"      "1250"       "3000"
## [16] "4000"       "22500"      "Refused"    "5250"       "2916"
## [21] "17500"      "45000"      "1750"       "11250"      "37500"
## [26] "4582"       "2500"       "35000"      "124950"     "75000"
## [31] "120000"     "65000"      "27500"      "22908"      "208"
## [36] "14578"      "1350000"    "125000"     "140000"     "175"
```

```
## [41] "166600"     "50000"      "225000"     "900000"     "1950000"
## [46] "18743"      "625"        "146"        "1874"       "70000"
## [51] "29155"      "55000"      "104"        "750000"     "85000"
## [56] "52500"      "1041"       "30000"      "5415"       "3749"
## [61] "7081"       "675000"     "90000"      "10413"      "2000000"
## [66] "1458"       "333"        "37485"      "525000"     "340000"
## [71] "450000"     "62475"      "275000"     "175000"     "2291"
## [76] "350000"     "50"         "180000"     "1500000"    "110000"
## [81] "375000"
```

```
finscope_data_useful$BusTurnMonthly[finscope_data_useful$BusTurnMonthly ==
"refused"] <- NA
finscope_data_useful$BusTurnMonthly[finscope_data_useful$BusTurnMonthly ==
"Don't know"] <- NA
finscope_data_useful$BusTurnMonthly <-
as.numeric(finscope_data_useful$BusTurnMonthly,na.rm = TRUE)
```

```
## Warning: NAs introduced by coercion
```

```
summary(finscope_data_useful$BusTurnMonthly)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##      50    1500    5000   26456   12500 2000000    2898
```

```
#Monthly profits..
unique(finscope_data_useful$BusNetProfmnthly)
```

```
##  [1] "4000"       "3000"       "1750"       "1500"       "Don't know"
##  [6] "200"        "175"        "375"        "700"        "1250"
## [11] "5000"       "Refused"    "22500"      "5250"       "1041"
## [16] "7500"       "12500"      "750"        "0"          "2500"
## [21] "11250"      "10413"      "45000"      "17500"      "7000"
## [26] "14578"      "10000"      "62"         "225000"     "625"
## [31] "52500"      "22908"      "27500"      "37500"      "16000"
## [36] "50"         "1874"       "333"        "75000"      "35000"
## [41] "2916"       "104"        "4582"       "7081"       "3749"
## [46] "2291"       "30000"      "1458"       "31"         "208"
## [51] "350000"     "146"        "124950"     "166600"     "29155"
## [56] "2000000"    "55000"      "65000"      "5415"       "125000"
## [61] "175000"     "180000"     "120000"     "62475"      "37485"
## [66] "50000"      "1500000"    "1350000"    "85000"      "18743"
## [71] "450000"     "750000"     "375000"
```

```
finscope_data_useful$BusNetProfmnthly[finscope_data_useful$BusNetProfmnthly
== "refused"] <- NA
finscope_data_useful$BusNetProfmnthly[finscope_data_useful$BusNetProfmnthly
== "Don't know"] <- NA
finscope_data_useful$BusNetProfmnthly <-
as.numeric(finscope_data_useful$BusNetProfmnthly,na.rm = TRUE)
```

```
## Warning: NAs introduced by coercion
```

```
summary(finscope_data_useful$BusNetProfmnthly)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0     750    1500    9964    5250 2000000    2904
```

Have Access to how many business functions.

Then the second variable we create is the business functions variable, which tells us how many business functions each SMME has access to. This variable can assist in shedding light on the sophistication of each of these SMMEs, and provide more information that might be relevant in drawing distinctions between the enterprises that are more profitable and resilient from those that are not. For example, the fact that a business has access to a business plan, marketing plan, or access to financial recording system and other facilities might shed light on whether an enterprise is professional in its business conduct, has a concrete plan on how it intends to take advantage of market opportunities and generate profits for its owners. Generally, it is reasonable to expect that the more business functions a business has access to, the more successful the business will be in executing its operations and thus the more profitable it will be. Properly documented business strategies with regard to the vision the business has and how it intends to generate value and grow are important in guiding the business on its path to achieving its most ambitious goals.

The creation of *HaveAccesstohowmanybusfuncs* variable is one of the most extensive variable creation schemes we will do in this analysis, as it requires that we use multiple functions to create one variable. Then, using the now created *finscope_data_useful* data frame, we use the pipe operator *%>%* to feed the data forward to the *mutate* function which is used to change the form of the responses in the respective variables from being characters such as *YES or NO* to being either *1 or 0*. We do this for all the variables pertaining to business information, with *YES* corresponding to *1* and *NO* corresponding to a *0*. Then, change their data types from being *character* to *numerical* data types using *as.numeric*. Then, we horizontally sum across the rows, where if there is *1* it adds up, and if it a *0* it remains the same. Finally, each row has the potential of having access to at least *6* business functions depending on how many business functions each of these businesses has access to. Note, there is order in the number of business functions each enterprise has access to. With the more access it has, the better its chances of being profitable.

```
#To create the new variable, we make extensive use of the mutate function
where we mutate each of the variables that have something to do with business
information from being encoded as a character of either YES/NO to being
encoded as either 1 or 0.

finscope_data_useful <- finscope_data_useful %>% mutate(`vision/mis` =
replace(`vision/mis`,`vision/mis` == "YES",1))
finscope_data_useful <- finscope_data_useful %>% mutate(`vision/mis` =
replace(`vision/mis`,`vision/mis` == "NO",0))
finscope_data_useful <- finscope_data_useful %>% mutate(busPln =
replace(busPln,busPln == "YES",1))
finscope_data_useful <- finscope_data_useful %>% mutate(busPln =
```

```r
replace(busPln,busPln == "NO",0))
finscope_data_useful <- finscope_data_useful %>% mutate(busstr =
replace(busstr,busstr == "YES",1))
finscope_data_useful <- finscope_data_useful %>% mutate(busstr =
replace(busstr,busstr == "NO",0))
finscope_data_useful <- finscope_data_useful %>% mutate(mrkpln =
replace(mrkpln,mrkpln == "YES",1))
finscope_data_useful <- finscope_data_useful %>% mutate(mrkpln =
replace(mrkpln,mrkpln == "NO",0))
finscope_data_useful <- finscope_data_useful %>% mutate(Accfinrec =
replace(Accfinrec,Accfinrec == "YES",1))
finscope_data_useful <- finscope_data_useful %>% mutate(Accfinrec =
replace(Accfinrec,Accfinrec == "NO",0))
finscope_data_useful <- finscope_data_useful %>% mutate(frmlTrnStff =
replace(frmlTrnStff,frmlTrnStff == "YES",1))
finscope_data_useful <- finscope_data_useful %>% mutate(frmlTrnStff =
replace(frmlTrnStff,frmlTrnStff == "NO",0))
finscope_data_useful <- finscope_data_useful %>% mutate(busbgt =
replace(busbgt,busbgt == "YES",1))
finscope_data_useful <- finscope_data_useful %>% mutate(busbgt =
replace(busbgt,busbgt == "NO",0))

#turning the columns into numeric variables which will be useful in the next
step/ we do this for each of the variables on business information.
finscope_data_useful$`vision/mis` <-
as.numeric(finscope_data_useful$`vision/mis`)
finscope_data_useful$busPln       <- as.numeric(finscope_data_useful$busPln)
finscope_data_useful$busstr       <- as.numeric(finscope_data_useful$busstr)
finscope_data_useful$mrkpln       <- as.numeric(finscope_data_useful$mrkpln)
finscope_data_useful$Accfinrec    <-
as.numeric(finscope_data_useful$Accfinrec)
finscope_data_useful$frmlTrnStff  <-
as.numeric(finscope_data_useful$frmlTrnStff)
finscope_data_useful$busbgt       <- as.numeric(finscope_data_useful$busbgt)

#Have access to how many business functions??/ from 0 to 6 business
functions/ thus,each business will have access to from 0 to 6 business
functions, we do this by adding up all the columns. This analysis will expect
that as the number of business functions a business has access to, the more
profitable/successful it'll be.
#Note, we have also made use of the pipe operator %>% to feed the dataset to
the other functions, and we also create a new variable called
HaveAccessToHowManyBusFuncs which store to the finscope_data_useful.
finscope_data_useful$HaveAccesstohowmanybusfuncs <- finscope_data_useful%>%
  select(`vision/mis`,busPln,busstr,mrkpln,Accfinrec,frmlTrnStff,busbgt)%>%
  rowSums(na.rm = TRUE)

summary(finscope_data_useful$HaveAccesstohowmanybusfuncs)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.2812  0.0000  7.0000
```

The next variable we will be cleaning is the categorical variable of provinces. It is worth noting that this is a categorical variable, not simply a binary category, thus, one of the levels in this variable will be used as the reference category. Thus, when we interpret each category, it will be interpreted in reference to the selected category.

With this variable, it is expected that the businesses that are found in the core economic/industrial provinces in South Africa will tend to perform better compared to those that are outside the industrial hubs. Thus, this analysis expects that businesses found in Gauteng, followed by the Western Cape, then KwaZulu-Natal will be significantly more positively associated with business performance than those found in the rest of the provinces which are not in the industrial hubs.

This variable will be encoded as a factor/categorical variable with nine levels. We have used *unique* to get a sense of what these categories will be if there will be missing values or values that we didn't expect so that we can be able to deal with such. Then, the *class* function was used to review the class these variables were stored so that if they were not appropriate they could be fixed. Then we used the *factor* function to encode the variables into nine categories and one variable will be used as the reference category. Note, here *R* simply organized the provinces in alphabetical order as this is the default way to organize variables in *R*. Then *summary* simply extracts the most salient features of the variables.

Note, that the way we will deal with categorical variables across this analysis will be the same. Thus for each variable below that is categorical, reference this variable to get an idea of the usage of the functions to deal with factor variables.

```
#Province/ the labels must correspond to the levels
unique(finscope_data_useful$Province)

## [1] "E.Cape"     "Gauteng"    "Free State" "W.Cape"     "KZN"
## [6] "Limpopo"    "Mpumalanga" "N.West"     "N.Cape"

class(finscope_data_useful$Province)

## [1] "character"

finscope_data_useful$Province <-factor(finscope_data_useful$Province)
summary(finscope_data_useful$Province)

##     E.Cape Free State    Gauteng        KZN    Limpopo Mpumalanga
N.Cape
##        727        445       1137        892        535        472
283
##     N.West     W.Cape
##        496        689

str(finscope_data_useful$Province)

##  Factor w/ 9 levels "E.Cape","Free State",..: 1 1 1 1 1 1 1 1 1 ...
```

Next up, we clean the location column, which has four levels. This variable will separate businesses according to the locations in which they are found. There are four main locations into which this variable separates into, which are rural formal, tribal area, urban formal, and urban informal. We expect that urban formal to have the most significant and positive relation with business performance, followed by Tribal formal then urban informal then, the most disadvantaged being tribal area. This would especially be the case in the instance of South Africa where the distinction between formal and informal would highly likely follow the reminiscences of apartheid-era spatial and economic planning which persist to this day in the division between the formal and the informal areas.

The logic for us to expect such a relationship is due to infrastructural and economic considerations. We expect businesses in urban areas to have better infrastructure which can aid in ensuring that they can deliver superior goods or services compared to the relatively infrastructural disadvantaged areas in the rural and informal areas. Infrastructure such as places to run business, being in the CBD where there's an influx of many people who are coming through in droves to buy goods or services can prove to be an advantage compared to the rural or informal sectors whose businesses might not be located in areas of commerce. Furthermore, in the context of South Africa, there are vast differences between tribal areas and rural formal as the latter is much more likely to be engaged in much more skilled and capital-intensive businesses compared to their more tribal rural area counterparts.

The second advantage that location offers pertains to economic advantages. Generally, people who are found in rural or urban informal areas tend to be relatively poorer compared to their urban and rural formal counterparts. Also, these two groups will tend to be located away from the economic core areas which have many buyers and sellers. This relative economic advantage these locations and inhabitants enjoy can translate to the businesses that operate in these areas having many more customers who can also spend relatively way more money and thus translate to better business performance compared to their poorer rural and urban informal counterparts.

Further, it is generally the case that market entrants will typically sell products to people or customers with whom they have a personal relationship with to facilitate trust in that transaction. Now, since people who are found in urban informal, and rural areas tend to be relatively poor, the basket of goods or services you can offer them is quite small as they are constrained by budget considerations, and most of their needs are already covered by the core economy. This limitation might be a bit relaxed for their relatively well-off counterparts who can afford to spend more on new products.

The second variable also aids the above, in helping to determine business performance. Unlike the above, this variable looks into where each SMME is operating within the respective locations. Some are operating at residential premises, whilst some are in more formal places that were tailor-designed for businesses, and others would be operating in say markets, school cafeterias, and many more. Again, we expect the more business premises are formal and tailor-made for that particular business, the more they will be able to drive profitability, compared to the enterprises whose premises were not designed for the particular business.

The construction of this variable also follows the same pattern we used to construct the previous variable above.

```
#Location/
unique(finscope_data_useful$Location)

## [1] "Urban formal"   "Urban informal" "Tribal area"    "Rural formal"

finscope_data_useful$Location <- factor(finscope_data_useful$Location)
class(finscope_data_useful$Location)

## [1] "factor"

summary(finscope_data_useful$Location)

##   Rural formal    Tribal area   Urban formal Urban informal
##            445           1370           3389            472

#WhereBusinessOperate
unique(finscope_data_useful$wherebusoperate)

##  [1] "Residential premises - dwelling/garage/building on residential
premises"
##  [2] "School Cafeteria"
##  [3] "Street/street corner/pavement"
##  [4] "Door to door/Go to customers"
##  [5] "Office block/office park"
##  [6] "Stall/table/container in a designated trading or market area"
##  [7] "Car/truck/vehicle"
##  [8] "Business park/Premises dedicated to my business -
hotel/accommodation facility/factory/workshop"
##  [9] "Shopping mall"
## [10] "Open space-Isipingo"
## [11] "Farm/small holding"
## [12] "Online - internet, phone selling"

finscope_data_useful$wherebusoperate <-
factor(finscope_data_useful$wherebusoperate)
levels(finscope_data_useful$wherebusoperate)

##  [1] "Business park/Premises dedicated to my business -
hotel/accommodation facility/factory/workshop"
##  [2] "Car/truck/vehicle"
##  [3] "Door to door/Go to customers"
##  [4] "Farm/small holding"
##  [5] "Office block/office park"
##  [6] "Online - internet, phone selling"
##  [7] "Open space-Isipingo"
##  [8] "Residential premises - dwelling/garage/building on residential
premises"
##  [9] "School Cafeteria"
## [10] "Shopping mall"
```

```
## [11] "Stall/table/container in a designated trading or market area"
## [12] "Street/street corner/pavement"

summary(finscope_data_useful$wherebusoperate)

## Business park/Premises dedicated to my business - hotel/accommodation
facility/factory/workshop
##
203
##
Car/truck/vehicle
##
34
##                                                                    Door to
door/Go to customers
##
223
##
Farm/small holding
##
103
##
Office block/office park
##
65
##                                                              Online -
internet, phone selling
##
14
##
Open space-Isipingo
##
1
##                          Residential premises - dwelling/garage/building on
residential premises
##
4017
##
School Cafeteria
##
39
##
Shopping mall
##
64
##                                    Stall/table/container in a designated
trading or market area
##
126
##
```

```
Street/street corner/pavement
##
787
```

Business type

Next, we look at the type of businesses each of these enterprises engage in. It is generally the case that different businesses in different industries will not perform the same. This can be attributed to a wide variety of factors, for example, some businesses are just more sophisticated and are able to serve the core needs of society compared to others. Others might have been started with little market research and thus might not be really solving a problem in society, and the intentions of the business founder were based on survivalist intents rather than thorough research into which opportunities can generate profitability. Thus, it will be worthwhile to figure out which types of business have a robust relationship with business profitability.

Here again, we didn't specify the levels, and *R* automatically orders the variable in alphabetical order which is desirable as this reduces the chance of making mistakes in labeling since labels have to match the specified levels. This might not be clear now, but if you play around with the *factor* function, you can see that you can tweak the specifications of this function.

```
#The types of businesses each of these enterprises engage in.
finscope_data_useful$Businesstype <-
factor(finscope_data_useful$Businesstype)
str(finscope_data_useful$Businesstype)

##  Factor w/ 13 levels "Grow something and sell, e.g. fruit, vegetables,
plants (like a nursery)",..: 9 9 9 9 9 9 9 8 9 9 ...

levels(finscope_data_useful$Businesstype)

##  [1] "Grow something and sell, e.g. fruit, vegetables, plants (like a
nursery)"
##  [2] "Rear livestock/poultry and sell e.g. chickens"
##  [3] "Render a professional service e.g. doctor, lawyer, accountant,
engineer, consultant"
##  [4] "Render a skilled service e.g. mechanic, plumber, hair salon, barber,
painting, landscaping"
##  [5] "Render building/construction services"
##  [6] "Render other services e.g. car wash, garden services, transport
(taxi services), catering"
##  [7] "Render tourism-related services e.g. accommodation/hotel/B&B/guest
house, tour operators"
##  [8] "Sell by-products of animals e.g. meat, eggs, milk"
##  [9] "Sell something in the same form that I buy from someone else (dont
add value, e.g. cigarettes)"
## [10] "Sell something that I buy but add value to, e.g. repackage, cook,
etc"
## [11] "Sell something that I collect from nature, e.g. herbs, firewood,
```

```
charcoal, thatch, sand, stone"
## [12] "Sell something that I get for free, e.g. second hand clothes, scrap
metal"
## [13] "Sell something that I make e.g. crafts, clothes, furniture, bricks"

summary(finscope_data_useful$Businesstype)

##                        Grow something and sell, e.g. fruit, vegetables,
plants (like a nursery)
##
242
##                                                  Rear livestock/poultry
and sell e.g. chickens
##
146
##            Render a professional service e.g. doctor, lawyer, accountant,
engineer, consultant
##
110
##     Render a skilled service e.g. mechanic, plumber, hair salon, barber,
painting, landscaping
##
917
##                                                          Render
building/construction services
##
90
##      Render other services e.g. car wash, garden services, transport (taxi
services), catering
##
400
##       Render tourism-related services e.g. accommodation/hotel/B&B/guest
house, tour operators
##
82
##                                               Sell by-products of animals
e.g. meat, eggs, milk
##
119
## Sell something in the same form that I buy from someone else (dont add
value, e.g. cigarettes)
##
2497
##                           Sell something that I buy but add value to, e.g.
repackage, cook, etc
##
633
## Sell something that I collect from nature, e.g. herbs, firewood, charcoal,
thatch, sand, stone
##
```

```
74
##                       Sell something that I get for free, e.g. second hand
clothes, scrap metal
##
56
##                                Sell something that I make e.g. crafts,
clothes, furniture, bricks
##
310
```

Own or Rent private residences and Own or rent business premises

The inclusion of the two variables below, which are owning or renting the private residence and business premises, plays a multifold role in the analysis of business performance. Let's first look at it from the perspective of property ownership as collateral. Collateral serves as a costly way for good borrowers to signal themselves by pledging a highly marketable good to secure a loan. The presence of such a good can serve as a measure to mitigate against risky behavior by the borrowers once they receive the credit. Property can act as a good proxy for collateral as property is highly immutable, meaning the owners cannot move it or exchange it easily. Further, we expect that property will retain its value for long sustained periods of time in relation to other assets such as machinery or vehicles.

Furthermore, the property is relatively more liquid compared to other assets and can be easily sold off. Thus, we expect that property will act as a good proxy for collateral. This then translates into the thinking that companies with properties can pledge them as collateral. Thus, it is highly likely that these businesses will in turn have access to credit which will aid in the growth of such businesses, compared to those with no collateral position.

Another relationship that property might have to business performance might have to do with having business premises. Businesses that can operate in their own business premises will be more secure and have reliable places to store their inventory or machinery or any other assets that are essential in the smooth running of the business. Furthermore, business premises can also offer a place where customers can easily reach these businesses and source such goods and further offer trust between the customer and the store owners as businesses with physical addresses will tend to be much more trustworthy.

Now, let us look at the cleaning of these variables. The code below introduces something a bit different as it creates a new variable and the variable also has missing values. The first task is to encode the *not applicable* and *other* as missing values, that is, encode them to *NA*, and this is really an easy way of dealing with a potentially nuanced problem. However, domain knowledge would be needed to think through how could a business potentially not have business premises. Also in the instance of other, there could be so many other considerations that speculating what those could be, and then having to come up with one word that encapsulates all those reasons would present an enormous task, hence this analysis decided to relegate them to missing and then create a binary variable.

The creation of this variable also is an elaborate affair as we have to make use of conditional statements to either classify as a 1 or a 0 depending on the category a data point might fall into. Here, we assign *own* to 1 and everything else is relegated to 0, note, that we could have also assigned *not applicable* and *other* to 0 as well, however, that would have been an easy get-out-of-jail card. Here, it's also the first time we make use of the level and label statements that I alluded to above, however, they are self-explanatory. One thing to need to be careful of is to make sure that the levels match the labels as *R* cannot help us on this one and there are high chances of making errors and thus collapsing the entire operation. The same reasoning and logic are also applicable to the second variable which looks into the ownership of private residences instead of business residences.

```
#Own or rent business premises
unique(finscope_data_useful$ownrent)

## [1] "Own"              "Use it (no rent)" "Not applicable"   "Rent"
## [5] "Other"

finscope_data_useful$ownrent[finscope_data_useful$ownrent == "Not
applicable"] <- NA
finscope_data_useful$ownrent[finscope_data_useful$ownrent == "Other"] <- NA
finscope_data_useful$ownrent <- ifelse(finscope_data_selected$ownrent ==
"Own",1,0)
finscope_data_useful$ownrent <- factor(finscope_data_useful$ownrent, levels =
c(0,1), labels = c("Don't own the business premises (Use it without rent/or
rent it)","Own the business premises"))
str(finscope_data_useful$ownrent)

##  Factor w/ 2 levels "Don't own the business premises (Use it without
rent/or rent it)",..: 2 2 2 2 2 2 2 1 2 ...

levels(finscope_data_useful$ownrent)

## [1] "Don't own the business premises (Use it without rent/or rent it)"
## [2] "Own the business premises"

summary(finscope_data_useful$ownrent)

## Don't own the business premises (Use it without rent/or rent it)
##                                                            2292
##                                        Own the business premises
##                                                            3384

#Own or rent private residences/ not the business facilities....
unique(finscope_data_useful$OwnerRentorOwnPrivRes)

## [1] "Own"               "Rent"              "Not Applicable"
## [4] "Stay without rent" "Other"

finscope_data_useful$OwnerRentorOwnPrivRes[finscope_data_useful$OwnerRentorOw
nPrivRes == "Not applicable"] <- NA
finscope_data_useful$OwnerRentorOwnPrivRes[finscope_data_useful$OwnerRentorOw
```

```
nPrivRes == "Other"] <- NA
finscope_data_useful$OwnerRentorOwnPrivRes <-
ifelse(finscope_data_selected$OwnerRentorOwnPrivRes == "Own",1,0)
finscope_data_useful$OwnerRentorOwnPrivRes <-
factor(finscope_data_useful$OwnerRentorOwnPrivRes, levels = c(0,1), labels =
c("Don't own the private residence (Use it without rent/or rent it)","Own the
private residence"))
summary(finscope_data_useful$OwnerRentorOwnPrivRes)

## Don't own the private residence (Use it without rent/or rent it)
##                                                            1052
##                                        Own the private residence
##                                                            4624
```

SMME Classification, number of employees, and Number of hours worked

The next variable we look into pertains to the separation of these enterprises into either Own account, Small, Micro, or Medium Enterprises (SMMEs classification).

The separation of these enterprises into these categories follows the logic that there are distinctions between larger firms and relatively smaller firms and the challenges they face and their performances tend to vary in accordance with their sizes. For instance, smaller enterprises will tend to be relatively more informal compared to their larger counterparts. Furthermore, larger enterprise owners will tend to have higher skill profiles or access to a budget to source such skills which can translate to better performance when compared to their smaller peers. Also, the opportunities they pursue in the markets and the sophistication in their deliverance of goods and services in response to such opportunities might differ significantly, with more efficiency attributed to the larger enterprises. This would be also the case when it comes to budgets, access to capital, or credit lines between enterprises of different levels.

Furthermore, the very small enterprises will tend to be survivalist in their nature and thus would be established to merely make ends meet for the owners and not to take advantage of business opportunities. Thus, these enterprises might approach the market with a scarcity mindset which does not look to take advantage of market opportunities but merely generate income to sustain the owner.

These points above seek to illustrate that these enterprises exist on a continuum and thus cannot be treated in the same way and they each have a unique and complex relationship with business performance.

Now, let us look at the construction of this variable. First, in our analysis, we used a nested if-else statement to classify these different enterprises according to the number of employees and then classified these enterprises into either of these four classes/categories, own account if there are 0 workers, micro-enterprises if it had fewer than 10 enterprises, a small enterprise if it had less than 49 employees and the last category was the medium enterprise category. Now, since these categories are not arbitrary and there is some order to them, we set *ordered* to *TRUE* and set the levels in their respective order as guided by the number of employees.

The next variable is related to the above variable, here we just putting it as it is without the manipulations. Now, this might pose problems later on since it's perfectly related to the enterprise classification variable, thus including them both in a model might pose issues. We will sort it out later on.

We also have the number of hours worked. We expect that the more hours worked, the more productivity and thus the more profits generated. However, this needs not be the case as some businesses are more advanced and thus might generate more profitability even with fewer hours worked.

```
#Total no of workers/enterprise classification..
finscope_data_useful$EnterpriseClassification <-
finscope_data_useful$TotNumWorkers
finscope_data_useful$EnterpriseClassification <-
ifelse(finscope_data_useful$EnterpriseClassification > 49,"Medium
Enterprise",ifelse(finscope_data_useful$EnterpriseClassification  > 10,"Small
Enterprise",ifelse(finscope_data_useful$EnterpriseClassification>0,"Micro
enterprise","Own Account")))
finscope_data_useful$EnterpriseClassification <-
factor(finscope_data_useful$EnterpriseClassification,ordered =  TRUE, levels
= c("Own Account","Micro enterprise","Small Enterprise","Medium Enterprise"))
class(finscope_data_useful$EnterpriseClassification)

## [1] "ordered" "factor"

str(finscope_data_useful$EnterpriseClassification)

##  Ord.factor w/ 4 levels "Own Account"<..: 2 2 1 1 2 2 2 2 1 1 ...

summary(finscope_data_useful$EnterpriseClassification)

##       Own Account  Micro enterprise  Small Enterprise Medium Enterprise
##              3715              1857                94                10

#Check the number of workers..
unique(finscope_data_useful$TotNumWorkers)

##  [1]   3   2   0   1   4   6   7   5  10   9  13  11  12   8  23  18  45
15  50
## [20]  14  34  40  42  24  20  21  27  79 198  25  97  22 150  17  19  16
60  30
## [39]  47  48 107  33  35  37 200  32  31

class(finscope_data_useful$TotNumWorkers)

## [1] "numeric"

summary(finscope_data_useful$TotNumWorkers)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   1.335   1.000 200.000
```

```
#total hours worked
class(finscope_data_useful$TotHoursWrk)

## [1] "numeric"

summary(finscope_data_useful$TotHoursWrk)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   8.000   9.000   9.408  12.000  24.000       5
```

Education Level

Next, we deal with the education variable. This variable will also follow the patterns we have established above and the only major notable difference pertains to the fact that education like enterprise classification is also *ordered* and thus we set it to *TRUE.* The order for education is intuitive to follow as it starts with primary education and builds up all the way to post-matric which is characterized by various levels such as a university degree or other forms of post-matric training.

This variable will also offer value in our understanding of the factors that contribute to the success of enterprises. We generally expect businesses whose owners have a decent level of education, such as, at the least some high school education to contribute positively towards the growth of their businesses. The general assumption is that they have the adequate know-how to understand market opportunities, enter into contracts, and make use of the available business services such as banking, computerized business systems, loan applications, and many other factors relevant to the success of a business and are knowledge-based.

Further, we expect that education and training to be a good proxy for a manager's competence, understanding of the industry in which the business is operating, the market opportunities that exist in the market, and how to optimally allocate capital and production facilities to take advantage of such opportunities. This is very important in the South African context as small enterprises will typically face intense competition from the core economy which is well organized. Furthermore, we expect a significant portion of companies that have owners with high levels of education to operate businesses that are high-tech, knowledge and skills-based, which require extensive training which turn tend to generate robust profits due to their sophistication.

```
#Education level/ note, the education levels are ordered thus we ordered is
true
unique(finscope_data_selected$HighlevelEdu)

## [1] "Some high school"
## [2] "Matric"
## [3] "Some primary school"
## [4] "Primary school completed"
## [5] "No schooling"
## [6] "Post matric qualification (diploma)"
## [7] "University degree (undergrad/postgrad/masters/honours)"
## [8] "Apprenticeship"
```

```r
finscope_data_selected$Edu <- factor(finscope_data_selected$HighlevelEdu)
summary(finscope_data_selected$Edu)
```

```
##                                          Apprenticeship
##                                                      97
##                                                  Matric
##                                                    1650
##                                             No schooling
##                                                     148
##                     Post matric qualification (diploma)
##                                                     432
##                                 Primary school completed
##                                                     528
##                                        Some high school
##                                                    2207
##                                     Some primary school
##                                                     395
## University degree (undergrad/postgrad/masters/honours)
##                                                     219
```

```r
finscope_data_useful$ HighlevelEdu<-
factor(finscope_data_useful$HighlevelEdu,ordered =  TRUE, levels = c("No
schooling","Some primary school","Primary school completed","Some high
school","Matric","Apprenticeship","Post matric qualification
(diploma)","University degree (undergrad/postgrad/masters/honours)"))
summary(finscope_data_useful$HighlevelEdu)
```

```
##                                             No schooling
##                                                     148
##                                     Some primary school
##                                                     395
##                                 Primary school completed
##                                                     528
##                                        Some high school
##                                                    2207
##                                                  Matric
##                                                    1650
##                                          Apprenticeship
##                                                      97
##                     Post matric qualification (diploma)
##                                                     432
## University degree (undergrad/postgrad/masters/honours)
##                                                     219
```

```r
str(finscope_data_useful$HighlevelEdu)
```

```
##  Ord.factor w/ 8 levels "No schooling"<..: 4 4 4 5 4 4 5 2 4 2 ...
```

```r
unique(finscope_data_useful$HighlevelEdu)
```

```
## [1] Some high school
## [2] Matric
## [3] Some primary school
## [4] Primary school completed
## [5] No schooling
## [6] Post matric qualification (diploma)
## [7] University degree (undergrad/postgrad/masters/honours)
## [8] Apprenticeship
## 8 Levels: No schooling < Some primary school < ... < University degree
(undergrad/postgrad/masters/honours)
```

```
class(finscope_data_useful$HighlevelEdu)
```

```
## [1] "ordered" "factor"
```

To Whom the Business Sell to:

We now shift our attention to the variables that pertain to whom the business sells to. Here, we have a bunch of variables such as selling to the government, selling to other small enterprises, selling to larger enterprises, and selling to private individuals. These variables might offer insights pertaining to customers which are important in driving business growth. This analysis expects that the bigger the institutions the business sells to, the more reliable they will be as customers and thus will be important in determining profitability. For example, private individuals are expected to be less valuable as customers and to be more variable compared to much larger customers like other businesses in terms of the revenue they bring in.

This analysis expects that the following order will be followed in determining which customers are more important and thus reliable as sources of profitability. The order is expected to go as follows: larger enterprises, other small enterprises, government, and then private individuals. Note, that the size we alluded to above pertains to whether the customer is another business or a private individual, that is how big a customer is. This can be attributed to the fact that bigger customers will tend to have higher budgets compared to the smaller enterprises, and thus will be able to spend more. Furthermore, if a business sells to another business, it is highly likely that the good or service it sells to that respective entity is critical to the survival of that particular business, or is an integral component to the functioning of that business that is buying.

Note, in the construction of these variables, in one instance we just went straight to the *labels* and we did not specify the *level* argument. For instance, in the construction of selling to government variable, we went straight to specifying the labels and we did not specify the level. Whereas with other variables such as the one for selling to private individuals, we also specified the *levels* of argument. Well, in the instance where we didn't specify the *level* argument, we made sure that the labeling matches the default levels set by *R* by following the alphabetical order of *Yes* or *No* which were the original labels. This variable construction is similar to the other variables we have dealt with before in this analysis, a slight difference is the fact that we have labeled this variable, and again ensuring that these labels correspond to the default levels set up by *R*.

```r
#Selling to whom variables...
#Selling to government variable
unique(finscope_data_useful$Gov)
```

```
## [1] "NO"  "YES"
```

```r
finscope_data_useful$Gov <- factor(finscope_data_useful$Gov, labels =
c("Doesn't sell to government","Sells to government"))
str(finscope_data_useful$Gov)
```

```
##  Factor w/ 2 levels "Doesn't sell to government",..: 1 1 1 1 1 1 1 1 1 1
...
```

```r
summary(finscope_data_useful$Gov)
```

```
## Doesn't sell to government        Sells to government
##                       5549                        127
```

```r
#To other small enterprise
unique(finscope_data_useful$otherSmallBus)
```

```
## [1] "NO"  "YES"
```

```r
finscope_data_useful$otherSmallBus <-
factor(finscope_data_useful$otherSmallBus,labels = c("Doesn't sell to other
small enterprises","Sells to other small enterpises"))
str(finscope_data_useful$otherSmallBus)
```

```
##  Factor w/ 2 levels "Doesn't sell to other small enterprises",..: 1 1 1 1
1 1 1 1 1 1 ...
```

```r
summary(finscope_data_useful$otherSmallBus)
```

```
## Doesn't sell to other small enterprises         Sells to other small
enterpises
##                                    4999
677
```

```r
#selling to private individuals
unique(finscope_data_selected$PrivIndv)
```

```
## [1] "YES" "NO"
```

```r
unique(finscope_data_useful$PrivIndv)
```

```
## [1] "YES" "NO"
```

```r
finscope_data_useful$PrivIndv <- factor(finscope_data_useful$PrivIndv,levels=
c("YES","NO"),labels = c("Sells to Private Individuals","Doesn't sell to
Private Individuals"))
levels(finscope_data_useful$PrivIndv)
```

```
## [1] "Sells to Private Individuals"        "Doesn't sell to Private
Individuals"
```

```
str(finscope_data_useful$PrivIndv)
```

```
##  Factor w/ 2 levels "Sells to Private Individuals",..: 1 1 1 1 1 1 1 1 1 1
...
```

```
summary(finscope_data_useful$PrivIndv)
```

```
##        Sells to Private Individuals Doesn't sell to Private Individuals
##                                5558                                  118
```

```
#Selling to other large business
unique(finscope_data_useful$OtherLargeBus)
```

```
## [1] "NO"  "YES"
```

```
finscope_data_useful$OtherLargeBus <-
factor(finscope_data_useful$OtherLargeBus,levels= c("NO","YES"), labels =
c("Doesn't sell to Larger enterprises","Sells to larger enterpises"))
str(finscope_data_useful$OtherLargeBus)
```

```
##  Factor w/ 2 levels "Doesn't sell to Larger enterprises",..: 1 1 1 1 1 1 1
1 1 1 ...
```

```
summary(finscope_data_useful$OtherLargeBus)
```

```
## Doesn't sell to Larger enterprises         Sells to larger enterpises
##                                5447                                229
```

Is Business Only source of income

Next, we look into the variable that probes whether a business is the only source of income or not. This variable contribution to our understanding of business performance is a bit complicated as it can be confounded by other factors. For example, households with own-account or micro-enterprises might also be recipients of social assistance grants, or the owners might be working part-time at other jobs since these enterprises' profitability might be limited, whereas the owners of the larger enterprises might have several businesses that provide alternative sources of income to these owners. However, the main relationship that is of major interest to this analysis is to see if reliance on a business as a source of income might contribute towards a positive business performance or not.

We expect that, though the confounding issue might cloud our judgment the model's ability to decipher its influence from the other confounding effect might not be strong. However, we expect that those businesses that serve as the owner's only source of income to be a bit more successful. We expect these owners to pour in a considerable amount of energy and resources into them and ensure their success, as they rely on them as a source of income. Further, the more an owner spends time on a business, we expect that this will translate to experience in the market and thus subsequently better performances.

```
#Is business only source of income
finscope_data_useful$`IsBusOnlySrcInc?` <-
factor(finscope_data_useful$`IsBusOnlySrcInc?`,labels = c("Business in not
```

```
the only source of income","Business is the only source of income"))
str(finscope_data_useful$`IsBusOnlySrcInc?`)

##  Factor w/ 2 levels "Business in not the only source of income",..: 2 2 1
2 1 2 2 2 2 1 ...

summary(finscope_data_useful$`IsBusOnlySrcInc?`)

## Business in not the only source of income
##                                      2005
##     Business is the only source of income
##                                      3671
```

Keeping Financial Records and Computerised records

Another variable we construct pertains to whether a business keeps financial records or not. Keeping financial financial is a very important aspect of a business and determining its performance. Keeping financial records can be a proxy for the accounting standards or lack thereof in the respective business. Proper records can ensure that the business is keeping track of the sources of its funds, is aware of its liabilities and assets at all times, and actively managing them. Good financial records also speak to the reliability of the data the business has to offer and whether we could reliably verify the business's profitability and performance. These records can also go a long way in ensuring that the business owners have a good understanding of their business at any point in time, and thus are aware that it is not leaking cash or are frivolous expenses that don't bring much value to the business, and at all times there is knowledge of the areas that this business is making profits. Further, these records are of value to the investors and creditors of the business who, with their capital can aid the business to grow in leaps and bounds.

The next variable relates to the previous variable we have dealt with, which looked at the presence or absence of financial records. Conceptually, this variable will be playing the same role as the previous one, in the sense that we will also be looking at whether a business keeps financial records or not. However, one major shift we expect will pertain to the quality of these financial records. Essentially, we expect the relationship between computerized financial records and business performance to be much more robust compared to financial records where it was not specified whether they are computerized or not. This thinking stems from the fact that computerized records will tend to be of generally higher quality and more standardized compared to other forms of recording that are not computerized. Furthermore, computerized financial records are easily transferable and less prone to error.

The construction of these variables follows the pattern of the other variables above. We also labeled the variables with a label that corresponds to the default R levels.

```
#Keep financial record/ awwww the default for the level is alphabetically...
unique(finscope_data_useful$KeepFinRec)

## [1] "Yes" "No"
```

```
finscope_data_useful$KeepFinRec <- factor(finscope_data_useful$KeepFinRec,
labels = c("Doesn't keep financial records","Keep financial records"))
str(finscope_data_useful$KeepFinRec)

##  Factor w/ 2 levels "Doesn't keep financial records",..: 2 2 1 2 1 1 2 1 1
1 ...

summary(finscope_data_useful$KeepFinRec)

## Doesn't keep financial records         Keep financial records
##                           2989                           2687

#Computirized financial records...
unique(finscope_data_useful$CompFinRecs)

## [1] "NO"  "YES"

finscope_data_useful$CompFinRecs <- factor(finscope_data_useful$CompFinRecs,
labels= c("Business does not keep computerized financial records","Business
keeps computerized financial records"))
str(finscope_data_useful$CompFinRecs)

##  Factor w/ 2 levels "Business does not keep computerized financial
records",..: 1 1 1 1 1 1 1 1 1 1 ...

summary(finscope_data_useful$CompFinRecs)

## Business does not keep computerized financial records
##                                                  5075
##         Business keeps computerized financial records
##                                                   601
```

Access to How many Insurance Products

The next variable pertains to how many business insurance products each enterprise has access to. The construction of this variable was also an elaborate affair and involved invoking many functions. However, as elaborate as it was, we will not explain the process here due to the fact that the construction of this variable is similar to the technique we used for the variable which probed whether a business has access to how many business functions.

The importance of this variable speaks to the risk attitudes of the business owners and whether they have risk mitigation measures in place. Logic dictates that those businesses with the proper risk measures will tend to be more successful and perform better. The reasoning is two-fold, one, generally if a business has valuable inventory or machinery, if they take measures to protect such assets, then, in the long run, would be more profitable as should anything happen to such assets they would be covered, and thus, this would not break the flow of the business operations due to an adverse event.

Another reason would stem from the viewpoint of investors or creditors. An investor would generally when coming to a business try to avoid the so-called principal-agent

problem, where the agent is the business if given a line of credit or investment capital would be reckless with such capital, and potentially act in a manner that is not consistent with the interests of the investors or creditors which are the principals. Thus, proper attitudes towards risk as demonstrated by taking out insurance can act as a costly signal to the investors or creditors that they are reliable and can extend the line of credit or investment. And, through the peculiar way we have constructed this variable, the more insurance products a small enterprise has, the more costly the signal they demonstrate to the market and thus the more worthy of investments or credit they are, thus again over the long run can attract capital and grow, leading to relatively better business performance.

The second variable which pertains to insurance also follows the same logic in its importance, however, it is set up to be a binary variable instead of a count variable. There might be an issue of perfect correlation with the constructed variable, but that will be taken care of when we do correlation analysis.

```
#Creating the insurance variable/ At least one form of business insurance...
unique(finscope_data_useful$BusContInsOffEquip)

## [1] "Never had"                      "Have now"
## [3] "Used to have but dont have now" "refused"

finscope_data_useful$BusContInsOffEquip[finscope_data_useful$BusContInsOffEqu
ip == "refused"] <- NA
finscope_data_useful <- finscope_data_useful %>% mutate(BusContInsOffEquip =
replace(BusContInsOffEquip,BusContInsOffEquip == "Have now",1))
finscope_data_useful <- finscope_data_useful %>% mutate(BusContInsOffEquip =
replace(BusContInsOffEquip,BusContInsOffEquip == "Never had",0))
finscope_data_useful <- finscope_data_useful %>% mutate(BusContInsOffEquip =
replace(BusContInsOffEquip,BusContInsOffEquip == "Used to have but dont have
now",0))
unique(finscope_data_useful$BusContSpecialisedToolsorMchinery)

## [1] "Never had"                      "Have now"
## [3] "Used to have but dont have now" "refused"

finscope_data_useful$BusContSpecialisedToolsorMchinery[finscope_data_useful$B
usContSpecialisedToolsorMchinery == "refused"] <- NA
finscope_data_useful <- finscope_data_useful %>%
mutate(BusContSpecialisedToolsorMchinery =
replace(BusContSpecialisedToolsorMchinery,BusContSpecialisedToolsorMchinery
== "Have now",1))
finscope_data_useful <- finscope_data_useful %>%
mutate(BusContSpecialisedToolsorMchinery =
replace(BusContSpecialisedToolsorMchinery,BusContSpecialisedToolsorMchinery
== "Never had",0))
finscope_data_useful <- finscope_data_useful %>%
mutate(BusContSpecialisedToolsorMchinery =
replace(BusContSpecialisedToolsorMchinery,BusContSpecialisedToolsorMchinery
== "Used to have but dont have now",0))
unique(finscope_data_useful$PrpStrBusPremIns)
```

```
## [1] "Never had"                       "Have now"
## [3] "Used to have but dont have now" "refused"

finscope_data_useful$PrpStrBusPremIns[finscope_data_useful$PrpStrBusPremIns
== "refused"] <- NA
finscope_data_useful <- finscope_data_useful %>% mutate(PrpStrBusPremIns =
replace(PrpStrBusPremIns,PrpStrBusPremIns == "Have now",1))
finscope_data_useful <- finscope_data_useful %>% mutate(PrpStrBusPremIns =
replace(PrpStrBusPremIns,PrpStrBusPremIns == "Never had",0))
finscope_data_useful <- finscope_data_useful %>% mutate(PrpStrBusPremIns =
replace(PrpStrBusPremIns,PrpStrBusPremIns == "Used to have but dont have
now",0))
unique(finscope_data_useful$CrpIns)

## [1] "Never had"                       "Used to have but dont have now"
## [3] "Have now"                        "refused"

finscope_data_useful$CrpIns[finscope_data_useful$CrpIns == "refused"] <- NA
finscope_data_useful <- finscope_data_useful %>% mutate(CrpIns =
replace(CrpIns,CrpIns == "Have now",1))
finscope_data_useful <- finscope_data_useful %>% mutate(CrpIns =
replace(CrpIns,CrpIns == "Never had",0))
finscope_data_useful <- finscope_data_useful %>% mutate(CrpIns =
replace(CrpIns,CrpIns == "Used to have but dont have now",0))
unique(finscope_data_useful$AccDamTransIns)

## [1] "Never had"                       "Have now"
## [3] "Used to have but dont have now" "refused"

finscope_data_useful$AccDamTransIns[finscope_data_useful$AccDamTransIns ==
"refused"] <- NA
finscope_data_useful <- finscope_data_useful %>% mutate(AccDamTransIns =
replace(AccDamTransIns,AccDamTransIns == "Have now",1))
finscope_data_useful <- finscope_data_useful %>% mutate(AccDamTransIns =
replace(AccDamTransIns,AccDamTransIns == "Never had",0))
finscope_data_useful <- finscope_data_useful %>% mutate(AccDamTransIns =
replace(AccDamTransIns,AccDamTransIns == "Used to have but dont have now",0))

#Turning into numeric
finscope_data_useful$BusContInsOffEquip <-
as.numeric(finscope_data_useful$BusContInsOffEquip)
finscope_data_useful$BusContSpecialisedToolsorMchinery <-
as.numeric(finscope_data_useful$BusContSpecialisedToolsorMchinery,na.rm =
TRUE)
finscope_data_useful$PrpStrBusPremIns <-
as.numeric(finscope_data_useful$PrpStrBusPremIns)
finscope_data_useful$CrpIns <- as.numeric(finscope_data_useful$CrpIns)
finscope_data_useful$AccDamTransIns <-
as.numeric(finscope_data_useful$AccDamTransIns)

#Summing up the rows
```

```r
finscope_data_useful$HaveAccesToHowmanyInsProd <- finscope_data_useful%>%

select(BusContInsOffEquip,BusContSpecialisedToolsorMchinery,PrpStrBusPremIns,
CrpIns,AccDamTransIns)%>%
  rowSums(na.rm = TRUE)
summary(finscope_data_useful$HaveAccesToHowmanyInsProd)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.1543  0.0000  5.0000

#Having insurance/Binary choice between having insurance and not having
insurance
unique(finscope_data_useful$HaveIns)

## [1] "Dont have" "Have"

finscope_data_useful$HaveIns <- factor(finscope_data_useful$HaveIns, labels =
c("Don't have insurance","Have Insurance"))
summary(finscope_data_useful$HaveIns)

## Don't have insurance       Have Insurance
##                 4205                 1471
```

Access to Financial Services, where a business get its credit and whether a business has received a business loan or not.

The next variable pertains to financial services access. The variable has four levels and is not ordered as there is no meaningful way to order them. The levels are as follows, banked, have access to formal financial services such as micro-finance institutions, have access to informal financial services like loan sharks or friends, and the last group is not served.

Again the importance of this variable pertains to the access to quality financial services which are essential in the smooth running of businesses. If a business is banked, this aids in establishing a structured and accurate financial trail for the business which helps it become attractive to potential investors or creditors. Also, access to these financial services can ensure that a business has secured ways of safe-keeping its cash which can go a long way in ensuring the sustainability of the business. Furthermore, the business through its established relationship with a financial services provider can gain access to other widely varying services besides banking, such as insurance products, lower interest rates, and help from these businesses to access newer markets.

Furthermore, there's a plethora of other indirect benefits from making use of financial services, such as more efficient payroll systems, pension funds for the employees, work injury compensation funds, and many other services that can boost staff morale, minimize the impact of work-related work injuries and litigation and many more benefits associated with having access to formal financial services. However, it is worth noting that this variable might not be able to capture in its entirety the complex nature of the financial services offered by businesses, and this catch-all variable might omit some important information pertaining to the quality and complexity of the services these enterprises have.

The variable that looks into where a business gets its credit from below is also related to the variable we worked on above. This one speaks directly to the source of credit for the enterprises surveyed, instead of financial inclusion in general. The source of credit for each of these enterprises reveals a lot about these enterprises and the quality of these businesses. We expect, like in the instance of financial access that businesses that borrow from the much more formal sources of credit like banks and the formal sector in general will have better performance compared to those that have more informal means of borrowing. Besides the fact that borrowing from more informal sources of credit is much more risky and expensive, we expect that if a company can be offered formal credit then that should say a lot about the quality and competence of such business, and thus is a great signal that the business performs well, is organized and has all the proper financial, asset trail in order. Thus, we expect that if a business has accessed credit from sources such as Banks or formal sources of credit then such business will have significant positive business performance.

```
#Probe into whether a business has access to financial services or not.
unique(finscope_data_useful$OvrallFinAccess)

## [1] "Banked"     "Not served" "Informal"   "Formal"

finscope_data_useful$OvrallFinAccess <-
factor(finscope_data_useful$OvrallFinAccess)
summary(finscope_data_useful$OvrallFinAccess)

##     Banked     Formal   Informal Not served
##       2782        204        335       2355

#Where does the business get credit from..
unique(finscope_data_useful$CreditBorrowingStrd)

## [1] "Not served"                    "Formal"
## [3] "Banked"                        "Informal"
## [5] "Borrowing from friends/family"

finscope_data_useful$SourceOfcredit <-
factor(finscope_data_useful$CreditBorrowingStrd)
summary(finscope_data_useful$SourceOfcredit)

##                         Banked Borrowing from friends/family
##                            661                            135
##                         Formal                       Informal
##                             16                             44
##                     Not served
##                           4820

#Variable on whether an enterprise received a bank loan or not\
unique(finscope_data_useful$BankLoan)

## [1] "NO"  "YES"

class(finscope_data_useful$BankLoan)
```

```
## [1] "character"

finscope_data_useful$BankLoan <- factor(finscope_data_useful$BankLoan)
summary(finscope_data_useful$BankLoan)

##   NO  YES
## 5537  139
```

Exporting products and importing supplies.

Another two variables of interest, are one which probes whether a business exports its products internationally and the one which looks at whether a business sources its supplies from international suppliers or not. Now these two variables have the potential of aiding us to establish the relative complexity and sophistication of the product offerings of such businesses.

This analysis assumes that a business for starters that has access to international markets will have the necessary expertise and competence in its product offering to compete in an international setting where it has to beat barriers such as regulation in different countries and different trade tariff systems. Furthermore, a business that has a presence in multiple markets is to a certain extent shielded from business fluctuations associated with the local markets, meaning they are diversified in their market access, and their profitability is not tied to one market and its whims. Thus, we expect that exporting to international markets should have a positive relationship with business performance.

The other variable has to do with where the business gets its supplies from. A business that can get suppliers from outside of its immediate geographical area signals that it possesses the necessary know-how to scour for resources from a wide range of sources. This might translate to efficiency in the ways in which the business renders its services or produces its goods as scouring for components and not accepting the most immediate suppliers signals resourcefulness. Also, searching for supplies from other countries can also aid in reducing costs of operations as importing can give access to the most competitive pricing from countries that have already developed the scale to develop such products. Thus, importing should be associated with positive business performance.

Now, the construction of these variables was similar to the other variables we have constructed above, thus we will spend no time in motivating their construction.

```
#Exporting
unique(finscope_data_useful$ExptoCUST)

## [1] "No"  "Yes"

finscope_data_useful$ExptoCUST <- factor(finscope_data_useful$ExptoCUST,
labels = c("Don't Export to outside of SA","Export to outside of SA"))
summary(finscope_data_useful$ExptoCUST)

## Don't Export to outside of SA       Export to outside of SA
##                          5633                            43
```

```
#Importing supplies/ having suppliers outside of SA
unique(finscope_data_useful$SuppOutofSA)

## [1] "No"  "Yes"

finscope_data_useful$SuppOutofSA <- factor(finscope_data_useful$SuppOutofSA,
labels = c("Don't have suppliers out of SA","Have suppliers out of SA"))
summary(finscope_data_useful$SuppOutofSA)

## Don't have suppliers out of SA       Have suppliers out of SA
##                           5641                             35
```

Does Business offer goods on Credit

Next, we look at the variable which looks into whether a business offers goods on credit or not. With this one, we expect that there would be a positive relationship between offering goods on credit and business performance. The logic for this goes as follows, if a business has the capacity to extend goods on credit, then this greatly expands the profit streams for the business. Though the risk of default also increases, if a business has a robust credit risk management system, offering goods on credit will tend to remove the constraints on the business from relying upon only customers who can pay now. However, the success of this strategy will be contingent on many confounding factors such as the nature of the products offered on credit, to whom these goods were offered, whether was it a government client private client or other business, and what mechanisms the business has to recuperate the credit and risk management policies in the business which informs the business's appetite for credit and which customers should be favored to be extended credit to.

The construction of this variable was however a straightforward affair. But, something that is a bit different with this variable is that the variable has three levels which is a deviation from the binary variables we usually work with.

```
#Offering goods on credit
unique(finscope_data_useful$OffGoodCred)

## [1] "No"             "Yes, sometimes" "Yes always"

finscope_data_useful$OffGoodCred <- finscope_data_selected$OffGoodCred
class(finscope_data_useful$OffGoodCred)

## [1] "character"

finscope_data_useful$OffGoodCred <-
factor(finscope_data_useful$OffGoodCred,labels = c("No","Yes always","Yes,
sometimes"))
levels(finscope_data_useful$OffGoodCred)

## [1] "No"             "Yes always"     "Yes, sometimes"

summary(finscope_data_useful$OffGoodCred)

##             No     Yes always Yes, sometimes
##           3657            480           1539
```

Number of Years Operating

Next, we look at the variable which looks into how many years a business has been operational. Again, we expect a positive relationship between businesses that have been around for many years and profitability. The more years a business has been operational, the more experience we assume that it will have in that particular market and thus be able to withstand those particular market fluctuations. Further, we assume that after many years in business, the business also has a clear view of what decisions and investments drive profitability and which don't. And, the fact that a business might have been business for so long, we expect that it should have built a sort of brand awareness in that market which translates to a customer base the business has serviced for years. All these factors, this analysis expects will translate into profitability and business success in the long term.

The second variable relates to the age of the business owner. We expect that the older the business owner is, the more experienced they are in terms of running their business, and thus the more profitable such businesses will tend to be.

The construction of this variable was different from the other variables we have since it was not given in the initial data set. First, we had to change the data class from being a character into a numeric variable, since a year is a numeric data type. Then, to get the age of each enterprise, we had to minus from 2010, the year in which this data-set was collected the variable *YearBusStart* which is the variable that stores the data when the business started. The difference between these two then gave us the number of years these businesses have been operational.

```
#Number of years business operational
class(finscope_data_useful$YearBusStart)

## [1] "character"

finscope_data_useful$YearBusStart <-
as.numeric(finscope_data_useful$YearBusStart)

## Warning: NAs introduced by coercion

summary(finscope_data_useful$YearBusStart)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1918    2001    2007    2004    2009    2010      78

finscope_data_useful$AgeOfBusiness <- 2010 -
finscope_data_useful$YearBusStart
summary(finscope_data_useful$AgeOfBusiness)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   1.000   3.000   6.109   9.000  92.000      78

#Age fix
class(finscope_data_useful$Age)

## [1] "character"
```

```r
finscope_data_useful$Age <- as.numeric(finscope_data_useful$Age,na.rm = TRUE)

## Warning: NAs introduced by coercion

summary(finscope_data_useful$Age)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   16.00   31.00   41.00   41.61   51.00   94.00      17

unique(finscope_data_useful$Age)

##  [1] 53 44 33 73 59 47 76 42 84 35 22 43 46 45 70 54 55 32 40 31 50 36 75
51 77
## [26] 49 57 30 60 56 72 23 26 25 NA 38 61 41 37 39 63 28 58 34 27 19 71 29
85 65
## [51] 52 20 79 74 64 24 21 48 68 18 69 16 62 17 66 67 80 87 81 82 78 89 94
88
```

Having Insurance Against Theft, Security measures and Whether has suffered crime 12 months prior.

Now we shift our attention to the variable that looks into whether a business has access to insurance against crime or not. This insurance product is also critical to assess, and it is especially the case in the South African context where crime in the literature has been cited as a major hindrance to business success due to the fact that SA has relatively high crime rates. Crime acts as a negative deterrence to business growth by dissuading business owners from investing in their businesses such as buying relatively more expensive business machinery due to fear that such assets might be at risk of being stolen or vandalized by criminals. This reluctance to invest in these businesses by the business owners translates into poor production choices being made on how to render services or produce products and thus leading to lower profitability. Furthermore, without insurance, even when the business has invested in such assets that bring about efficiency, if they get stolen, it might take longer to replace them leading to poor business performance or the owners might not be able to replace them at all. Thus, insurance, especially for the smaller enterprises that are disproportionally impacted by crime might mitigate the negative effects of crime. Thus, this translates to owners investing in their businesses with the hope that they would cover their losses should any adverse event occur.

The second variable also relates to crime and probes whether a business has security measures or not. This variable to some extent seeks to do the same job as the previous one. But, what it does differently is to check if the business has practical security measures besides insurance, such as burglar systems, alarm systems, or subscriptions to an armed response service. Again, this variable does not give an idea of how robust these security measures and it is highly likely that different enterprises depending on their level will have different approaches to security and its intensity. Note, the overall picture, besides the quality of security measures, is that businesses that have a proactive approach to crime will tend to fare better at withstanding its impact, and subsequently will be robust and be able to grow and generate profits.

The third variable also relates to crime and the logic that we have just outlined above also applies to it. However, this variable offers something a bit different and might help our model get a better sense of what informs business performance. Businesses that suffered from crime would probably have had their critical assets in disarray and thus be unable to perform as well as they should. This variable if there's enough variability in the data might shine light on the significance or lack thereof of crime affecting business performance after it has occurred. It could be that crime is not as much of an issue as we had initially thought.

```
#Have insurance against theft
unique(finscope_data_useful$ClaimInsforTheft)

## [1] "NO"  "YES"

finscope_data_useful$ClaimInsforTheft <-
factor(finscope_data_useful$ClaimInsforTheft, labels = c("Don't have
insurance claim against theft/crim","Have Insurance claim against crime or
theft"))
summary(finscope_data_useful$ClaimInsforTheft)

## Don't have insurance claim against theft/crim
##                                          5674
##   Have Insurance claim against crime or theft
##                                             2

#fix having security measures
finscope_data_useful$`HavingSecurityMeasures` <-
factor(finscope_data_useful$`HavingSecurityMeasures`,labels = c("Don't have
security measures","Have security Measures"))
str(finscope_data_useful$`HavingSecurityMeasures`)

##  Factor w/ 2 levels "Don't have security measures",..: 1 1 1 1 1 1 2 1 1 1
...

summary(finscope_data_useful$`HavingSecurityMeasures`)

## Don't have security measures       Have security Measures
##                         4745                          931

#fixingsufferedtheft/crime
unique(finscope_data_useful$sufferedCrimeortheft)

## [1] "No"  "Yes"

finscope_data_useful$sufferedCrimeortheft <-
factor(finscope_data_useful$sufferedCrimeortheft, labels= c("Business did not
suffer crime or theft in the last 12 months","Business suffered crime or
theft in the last 12 months"))
str(finscope_data_useful$sufferedCrimeortheft)

##  Factor w/ 2 levels "Business did not suffer crime or theft in the last 12
months",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
summary(finscope_data_useful$sufferedCrimeortheft)

## Business did not suffer crime or theft in the last 12 months
##                                                         5288
##       Business suffered crime or theft in the last 12 months
##                                                          388
```

Business Registration (Informality)

Now, this analysis shifts its focus towards whether a business is registered or not registered. This analysis will use this variable as a proxy for whether a business is an informal or a formal business. Why this distinction?, well it turns out this is very important in the context of South Africa where there's relatively a larger portion of the smaller enterprises being informal. The relatively smaller enterprises in South Africa can be divided into two, namely survivalist enterprises whose owners tend to be poorer, have a low skill profile and if given the opportunity would rather opt for opportunities in the formal sector. These types of enterprises will tend to be informal in their operations and lack the sophistication their larger counterparts have. Furthermore, these enterprises will also tend to generate relatively small incomes as they tend to go for business opportunities that usually have low barriers to entry and require little to no skill profile or start-up capital, thus they generally face stringent competition from the core economy.

On the flip side, still within these relatively small enterprises, you would find the so-called micro-enterprises or growth enterprises. These enterprises would tend to be a bit more formal in their approach to business tend to be more entrepreneurial driven and will respond to local opportunities for growth that exist in the market. Thus, these enterprises are much more motivated by profiteering instead of just going to business for survivalist reasons or livelihood. Therefore, these enterprises would tend to have higher prospects of growing into more profitable enterprises and being fully fledged formal enterprises. You'd typically find these businesses being your local tuck shops, welding businesses, businesses that sell alcohol, and many more.

Now, this distinction is really more important for the relatively smaller enterprises because its rarely the case that you going to find the relatively larger enterprises without being registered or being informal. Also, it is important to note that this variable is not a perfect proxy for formality as it might miss some growth enterprises that are not registered.

Again, I will not expand on the construction of this variable since it was already covered in similar variables.

```
#Business registration
finscope_data_useful$IsBusRegistered[finscope_data_useful$IsBusRegistered ==
"Dont know"] <- NA
unique(finscope_data_useful$IsBusRegistered)

## [1] "Yes" "No"  NA

finscope_data_useful$IsBusRegistered <-
factor(finscope_data_useful$IsBusRegistered, labels = c("Business is not
```

```
registered","Business is registered"))
str(finscope_data_useful$IsBusRegistered)

##  Factor w/ 2 levels "Business is not registered",..: 2 2 1 1 1 1 1 1 1 1
...

summary(finscope_data_useful$IsBusRegistered)

## Business is not registered     Business is registered
##                       4392                       1192
##                       NA's
##                         92
```

Making a Tender Submission.

The final two variables relate to whether these enterprises had before made a tender submission and if such a tender submission was successful or not. Again, this variable will provide us with some important insights when it comes to assessing business performance. First, it's reasonable to assume that if a business is in a state to submit a tender proposal, then that business is relatively well structured and has the expertise and competence to some extent to carry out the tasks outlined in the tender they are applying for. This is a great proxy for managerial competencies and in general the competence of the respective enterprise. This is the case due to the fact that if a company takes the initiative to apply for a tender, then that company and its leadership reasonably have an understanding of the industry it is operating.

The second variable looks into whether a firm's bid for a tender was successful or not. This second variable essentially enriches the first, as a company whose tender bid was successful validates that indeed that respective company is competent. Furthermore, a company that does not only identify lucrative market opportunities and bid for them but also wins such bids is a great company that will tend to be profitable. Further, having a successful tender application also bears testimony to the managers' competence, to identify good opportunities and execute them successfully is a great proxy for competence. Furthermore, successful tender applications translate into revenues for the business which contributes to profitability or good business performance.

```
#TenderSubmission
unique(finscope_data_useful$SubmittedTenderProp)

## [1] "No"  "Yes"

finscope_data_useful$SubmittedTenderProp<-
factor(finscope_data_useful$SubmittedTenderProp, labels = c("Did not submit a
tender application in the last 12 months","Submitted a tender application
last 12 months") )
str(finscope_data_useful$SubmittedTenderProp)

##  Factor w/ 2 levels "Did not submit a tender application in the last 12
months",..: 1 1 1 1 1 1 1 1 1 1 ...

summary(finscope_data_useful$SubmittedTenderProp)
```

```
## Did not submit a tender application in the last 12 months
##                                                      5479
##             Submitted a tender application last 12 months
##                                                       197
```

```
#TenderSucc
class(finscope_data_useful$TenderSucc)
```

```
## [1] "character"
```

```
unique(finscope_data_useful$TenderSucc)
```

```
## [1] "0"   "Yes" "No"
```

```
finscope_data_useful$TenderSucc[finscope_data_useful$TenderSucc == "0"] <-
"Did not apply"
unique(finscope_data_useful$TenderSucc)
```

```
## [1] "Did not apply" "Yes"           "No"
```

```
finscope_data_useful$TenderSucc<- factor(finscope_data_useful$TenderSucc)
str(finscope_data_useful$TenderSucc)
```

```
##  Factor w/ 3 levels "Did not apply",..: 1 1 1 1 1 1 1 1 1 1 ...
```

DATA AND COLUMN CLEANING:

Ensuring Variables have consistent names

First, we look at the dataframe we are working with to make sure that all the variables are of the form/datatype we expect them to be and that there are no variables that we missed when we were working with the dataset. From the code below, it seems like the most salient variables are encoded appropriately, and thus, we move forward with our analysis.

```
#checking for data types/ seems all the data types are of the desired type...
sapply(finscope_data_useful,class)
```

```
## $ID1
## [1] "numeric"
##
## $Location
## [1] "factor"
##
## $Province
## [1] "factor"
##
## $Businesstype
## [1] "factor"
##
## $wherebusoperate
## [1] "factor"
##
```

```
## $ownrent
## [1] "factor"
##
## $TotHoursWrk
## [1] "numeric"
##
## $`vision/mis`
## [1] "numeric"
##
## $busPln
## [1] "numeric"
##
## $busstr
## [1] "numeric"
##
## $mrkpln
## [1] "numeric"
##
## $Accfinrec
## [1] "numeric"
##
## $frmlTrnStff
## [1] "numeric"
##
## $busbgt
## [1] "numeric"
##
## $TotNumWorkers
## [1] "numeric"
##
## $PrivIndv
## [1] "factor"
##
## $otherSmallBus
## [1] "factor"
##
## $OtherLargeBus
## [1] "factor"
##
## $Gov
## [1] "factor"
##
## $Other
## [1] "character"
##
## $TenderSucc
## [1] "factor"
##
## $RegWthCIRPO
## [1] "character"
```

```
##
## $BusContInsOffEquip
## [1] "numeric"
##
## $BusContSpecialisedToolsorMchinery
## [1] "numeric"
##
## $PrpStrBusPremIns
## [1] "numeric"
##
## $CrpIns
## [1] "numeric"
##
## $AccDamTransIns
## [1] "numeric"
##
## $LargeSrcBorr
## [1] "character"
##
## $BankLoan
## [1] "factor"
##
## $BusTurnMonthly
## [1] "numeric"
##
## $BusNetProfmnthly
## [1] "numeric"
##
## $Age
## [1] "numeric"
##
## $Gender
## [1] "character"
##
## $Race
## [1] "character"
##
## $`Mar Stat`
## [1] "character"
##
## $HighlevelEdu
## [1] "ordered" "factor"
##
## $`IsBusOnlySrcInc?`
## [1] "factor"
##
## $KeepFinRec
## [1] "factor"
##
## $OvrallFinAccess
```

```
## [1] "factor"
##
## $CreditBorrowingStrd
## [1] "character"
##
## $HavingSecurityMeasures
## [1] "factor"
##
## $sufferedCrimeortheft
## [1] "factor"
##
## $ClaimInsforTheft
## [1] "factor"
##
## $IsBusRegistered
## [1] "factor"
##
## $SubmittedTenderProp
## [1] "factor"
##
## $BEEContrStatScoreCrd
## [1] "character"
##
## $CompFinRecs
## [1] "factor"
##
## $OwnerRentorOwnPrivRes
## [1] "factor"
##
## $ExptoCUST
## [1] "factor"
##
## $SuppOutofSA
## [1] "factor"
##
## $HaveIns
## [1] "factor"
##
## $OffGoodCred
## [1] "factor"
##
## $YearBusStart
## [1] "numeric"
##
## $HaveAccesstohowmanybusfuncs
## [1] "numeric"
##
## $EnterpriseClassification
## [1] "ordered" "factor"
##
```

```
## $HaveAccesToHowmanyInsProd
## [1] "numeric"
##
## $SourceOfcredit
## [1] "factor"
##
## $AgeOfBusiness
## [1] "numeric"
```

Next, we look whether there has been any duplicates in the process of naming the variables. We use the code *anyDuplicates* to do this , which checks the names using the *names* column to check if there are any names that are duplicates or not. And from the code below, there are zero duplicates from the results.

Then, we also use the pipe operator *$%>%* to pipe the data to the *remove_empty* function to remove any empty rows or columns.

```
#Seems we are all good in terms of duplicates/there's zero duplicated column
names.
anyDuplicated (names (finscope_data_useful))

## [1] 0

#Removes any empty rows or columns from the data.
finscope_data_useful <- finscope_data_useful %>% remove_empty(c("rows",
"cols"))

names(finscope_data_useful)

##  [1] "ID1"                                 "Location"
##  [3] "Province"                            "Businesstype"
##  [5] "wherebusoperate"                     "ownrent"
##  [7] "TotHoursWrk"                         "vision/mis"
##  [9] "busPln"                              "busstr"
## [11] "mrkpln"                              "Accfinrec"
## [13] "frmlTrnStff"                         "busbgt"
## [15] "TotNumWorkers"                       "PrivIndv"
## [17] "otherSmallBus"                       "OtherLargeBus"
## [19] "Gov"                                 "Other"
## [21] "TenderSucc"                          "RegWthCIRPO"
## [23] "BusContInsOffEquip"
"BusContSpecialisedToolsorMchinery"
## [25] "PrpStrBusPremIns"                    "CrpIns"
## [27] "AccDamTransIns"                      "LargeSrcBorr"
## [29] "BankLoan"                            "BusTurnMonthly"
## [31] "BusNetProfmnthly"                    "Age"
## [33] "Gender"                              "Race"
## [35] "Mar Stat"                            "HighlevelEdu"
## [37] "IsBusOnlySrcInc?"                    "KeepFinRec"
## [39] "OvrallFinAccess"                     "CreditBorrowingStrd"
## [41] "HavingSecurityMeasures"              "sufferedCrimeortheft"
```

```
## [43] "ClaimInsforTheft"                  "IsBusRegistered"
## [45] "SubmittedTenderProp"               "BEEContrStatScoreCrd"
## [47] "CompFinRecs"                       "OwnerRentorOwnPrivRes"
## [49] "ExptoCUST"                         "SuppOutofSA"
## [51] "HaveIns"                           "OffGoodCred"
## [53] "YearBusStart"                      "HaveAccesstohowmanybusfuncs"
## [55] "EnterpriseClassification"          "HaveAccesToHowmanyInsProd"
## [57] "SourceOfcredit"                    "AgeOfBusiness"
```

Now, we make sure that the column names are consistently structured and any characters that dirty up the column names are removed from the data. This function will make the column names more descriptive, remove any unnecessary spaces, and replace them with "_". For instance, *first name* would be cleaned into the much nicer version of *first_name* which is cleaner and consistent with the other values. And when we call the *names* function, we see that the names of the columns are nicely and consistently formatted.

```
#make the columns more consistent and descriptive
#Make the names more constitent and easier to work with/removes any
inconsistent characters that dirty up the variables/puts slash bars in
#between names, for example, will put replace first name with first_name..
finscope_data_useful <- finscope_data_useful %>% clean_names()
names(finscope_data_useful)
```

```
##  [1] "id1"
##  [2] "location"
##  [3] "province"
##  [4] "businesstype"
##  [5] "wherebusoperate"
##  [6] "ownrent"
##  [7] "tot_hours_wrk"
##  [8] "vision_mis"
##  [9] "bus_pln"
## [10] "busstr"
## [11] "mrkpln"
## [12] "accfinrec"
## [13] "frml_trn_stff"
## [14] "busbgt"
## [15] "tot_num_workers"
## [16] "priv_indv"
## [17] "other_small_bus"
## [18] "other_large_bus"
## [19] "gov"
## [20] "other"
## [21] "tender_succ"
## [22] "reg_wth_cirpo"
## [23] "bus_cont_ins_off_equip"
## [24] "bus_cont_specialised_toolsor_mchinery"
## [25] "prp_str_bus_prem_ins"
## [26] "crp_ins"
## [27] "acc_dam_trans_ins"
```

```
## [28] "large_src_borr"
## [29] "bank_loan"
## [30] "bus_turn_monthly"
## [31] "bus_net_profmnthly"
## [32] "age"
## [33] "gender"
## [34] "race"
## [35] "mar_stat"
## [36] "highlevel_edu"
## [37] "is_bus_only_src_inc"
## [38] "keep_fin_rec"
## [39] "ovrall_fin_access"
## [40] "credit_borrowing_strd"
## [41] "having_security_measures"
## [42] "suffered_crimeortheft"
## [43] "claim_insfor_theft"
## [44] "is_bus_registered"
## [45] "submitted_tender_prop"
## [46] "bee_contr_stat_score_crd"
## [47] "comp_fin_recs"
## [48] "owner_rentor_own_priv_res"
## [49] "expto_cust"
## [50] "supp_outof_sa"
## [51] "have_ins"
## [52] "off_good_cred"
## [53] "year_bus_start"
## [54] "have_accesstohowmanybusfuncs"
## [55] "enterprise_classification"
## [56] "have_acces_to_howmany_ins_prod"
## [57] "source_ofcredit"
## [58] "age_of_business"
```

Here, we ensure that all the missing values are encoded as *NA*, instead of all the other stuff that usually the missing bits are encoded as. To ensure we are able to appropriately deal with the missing bits in the data.

```
#convert values like "NA", "NULL", or "" to NA
finscope_data_useful <- finscope_data_useful %>% replace (., . == c ("NA",
"NULL", ""), NA)
```

Removing variables that don't have variability

Here, in this bit, we extract the most important features that we will use when we build our model, and we then pass them to a new dataframe called finscope_Useful_variables.

```
finscope_Useful_variables <- finscope_data_useful %>%
  select("location","province","businesstype","wherebusoperate","ownrent",
"tot_hours_wrk","tot_num_workers",
"priv_indv","other_small_bus","other_large_bus","gov",
"tender_succ","bank_loan", "bus_turn_monthly","bus_net_profmnthly","age",
"highlevel_edu","is_bus_only_src_inc","keep_fin_rec","ovrall_fin_access","hav
```

```
ing_security_measures","suffered_crimeortheft"
,"claim_insfor_theft","is_bus_registered","submitted_tender_prop"
,"comp_fin_recs","owner_rentor_own_priv_res","expto_cust","supp_outof_sa","ha
ve_ins","off_good_cred","have_accesstohowmanybusfuncs","enterprise_classifica
tion","have_acces_to_howmany_ins_prod","source_ofcredit","age_of_business")
```

Now, in this part of our analysis we are ensuring that we drop any unused levels, ensuring that each variable has levels that can also be found in the data set. Thus, we are essentially avoiding having variables with more levels than can be collaborated by the data. Thus, we will avoid the issue of having variables with only one variable, or too few observations per category.

```
sapply(finscope_Useful_variables, function(x) if(is.factor(x))
length(levels(x))) #Ensure that all the variables have the right number of
levels
```

```
## $location
## [1] 4
##
## $province
## [1] 9
##
## $businesstype
## [1] 13
##
## $wherebusoperate
## [1] 12
##
## $ownrent
## [1] 2
##
## $tot_hours_wrk
## NULL
##
## $tot_num_workers
## NULL
##
## $priv_indv
## [1] 2
##
## $other_small_bus
## [1] 2
##
## $other_large_bus
## [1] 2
##
## $gov
## [1] 2
##
## $tender_succ
```

```
## [1] 3
##
## $bank_loan
## [1] 2
##
## $bus_turn_monthly
## NULL
##
## $bus_net_profmnthly
## NULL
##
## $age
## NULL
##
## $highlevel_edu
## [1] 8
##
## $is_bus_only_src_inc
## [1] 2
##
## $keep_fin_rec
## [1] 2
##
## $ovrall_fin_access
## [1] 4
##
## $having_security_measures
## [1] 2
##
## $suffered_crimeortheft
## [1] 2
##
## $claim_insfor_theft
## [1] 2
##
## $is_bus_registered
## [1] 2
##
## $submitted_tender_prop
## [1] 2
##
## $comp_fin_recs
## [1] 2
##
## $owner_rentor_own_priv_res
## [1] 2
##
## $expto_cust
## [1] 2
##
```

```
## $supp_outof_sa
## [1] 2
##
## $have_ins
## [1] 2
##
## $off_good_cred
## [1] 3
##
## $have_accesstohowmanybusfuncs
## NULL
##
## $enterprise_classification
## [1] 4
##
## $have_acces_to_howmany_ins_prod
## NULL
##
## $source_ofcredit
## [1] 5
##
## $age_of_business
## NULL
```

```
finscope_Useful_variables <- droplevels(finscope_Useful_variables)
str(finscope_Useful_variables)
```

```
## tibble [5,676 × 36] (S3: tbl_df/tbl/data.frame)
##  $ location                      : Factor w/ 4 levels "Rural formal",..: 3
3 3 3 3 3 3 3 3 3 ...
##  $ province                      : Factor w/ 9 levels "E.Cape","Free
State",..: 1 1 1 1 1 1 1 1 1 ...
##  $ businesstype                  : Factor w/ 13 levels "Grow something and
sell, e.g. fruit, vegetables, plants (like a nursery)",..: 9 9 9 9 9 9 9 8 9
9 ...
##  $ wherebusoperate               : Factor w/ 12 levels "Business
park/Premises dedicated to my business - hotel/accommodation
facility/factory/workshop",..: 8 8 8 8 8 8 8 8 9 8 ...
##  $ ownrent                       : Factor w/ 2 levels "Don't own the
business premises (Use it without rent/or rent it)",..: 2 2 2 2 2 2 2 2 1 2
...
##  $ tot_hours_wrk                 : num [1:5676] 12 12 6 4 6 10 15 3 12 8
...
##  $ tot_num_workers               : num [1:5676] 3 2 0 0 1 1 1 1 0 0 ...
##  $ priv_indv                     : Factor w/ 2 levels "Sells to Private
Individuals",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ other_small_bus               : Factor w/ 2 levels "Doesn't sell to
other small enterprises",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ other_large_bus               : Factor w/ 2 levels "Doesn't sell to
Larger enterprises",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
##  $ gov                           : Factor w/ 2 levels "Doesn't sell to
government",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ tender_succ                   : Factor w/ 3 levels "Did not apply",..:
1 1 1 1 1 1 1 1 1 1 ...
##  $ bank_loan                     : Factor w/ 2 levels "NO","YES": 1 1 1 1
1 1 1 1 1 1 ...
##  $ bus_turn_monthly              : num [1:5676] 12500 220000 16000 7500
1500 7000 NA 1500 5000 200 ...
##  $ bus_net_profmnthly            : num [1:5676] 4000 3000 3000 1750 1500
1500 NA 1500 1500 200 ...
##  $ age                           : num [1:5676] 53 53 44 33 73 44 59 59 47
76 ...
##  $ highlevel_edu                 : Ord.factor w/ 8 levels "No
schooling"<..: 4 4 4 5 4 4 5 2 4 2 ...
##  $ is_bus_only_src_inc           : Factor w/ 2 levels "Business in not the
only source of income",..: 2 2 1 2 1 2 2 2 2 1 ...
##  $ keep_fin_rec                  : Factor w/ 2 levels "Doesn't keep
financial records",..: 2 2 1 2 1 1 2 1 1 1 ...
##  $ ovrall_fin_access             : Factor w/ 4 levels
"Banked","Formal",..: 1 1 4 1 1 1 3 4 3 4 ...
##  $ having_security_measures      : Factor w/ 2 levels "Don't have security
measures",..: 1 1 1 1 1 1 2 1 1 1 ...
##  $ suffered_crimeortheft         : Factor w/ 2 levels "Business did not
suffer crime or theft in the last 12 months",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ claim_insfor_theft            : Factor w/ 2 levels "Don't have
insurance claim against theft/crim",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ is_bus_registered             : Factor w/ 2 levels "Business is not
registered",..: 2 2 1 1 1 1 1 1 1 1 ...
##  $ submitted_tender_prop         : Factor w/ 2 levels "Did not submit a
tender application in the last 12 months",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ comp_fin_recs                 : Factor w/ 2 levels "Business does not
keep computerized financial records",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ owner_rentor_own_priv_res     : Factor w/ 2 levels "Don't own the
private residence (Use it without rent/or rent it)",..: 2 2 2 2 2 2 2 2 2 2
...
##  $ expto_cust                    : Factor w/ 2 levels "Don't Export to
outside of SA",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ supp_outof_sa                 : Factor w/ 2 levels "Don't have
suppliers out of SA",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ have_ins                      : Factor w/ 2 levels "Don't have
insurance",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ off_good_cred                 : Factor w/ 3 levels "No","Yes
always",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ have_accesstohowmanybusfuncs  : num [1:5676] 0 0 0 0 0 0 0 0 0 0 ...
##  $ enterprise_classification     : Ord.factor w/ 4 levels "Own
Account"<..: 2 2 1 1 2 2 2 2 1 1 ...
##  $ have_acces_to_howmany_ins_prod: num [1:5676] 0 0 0 0 0 0 0 0 0 0 ...
##  $ source_ofcredit               : Factor w/ 5 levels "Banked","Borrowing
from friends/family",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ age_of_business               : num [1:5676] 5 5 0 3 2 0 25 5 13 2 ...
```

From the above code, it should be that everything is fine, but when I try to run regression, we get the error that have a variable that has only one level. This issue might caused by a variable that has too few observations per category, but however, the above code can't pick it up because it is not an empty level per se, but rather has too few observations.

To solve this issue below, we will iteratively add one-factor variable at a time until we run into the error.

After this process, we find that the variable that checks for claiming insurance for theft is the one that causes the error and thus has to be reviewed.

```
#WORK ON THE DATA CLEANING ASPECT..
finscope_OutClean <- finscope_Useful_variables
#model1 <- lm(bus_net_profmnthly~province + enterprise_classification +
other_small_bus + gov + priv_indv + source_ofcredit + claim_insfor_theft
,data= finscope_OutClean)
#summary(model1)
```

In this analysis when we run regression on the data, we have had issues with the model as one of the variables keeps on causing errors and collapsing the model. Turns out that the variable that was collapsing the model was the one that looked into whether businesses claim insurance for theft, and one of the categories had only 2 observations thus this warranty we dropped it from the model. It was hard to be picked by the *droplevels* since it was not completely empty. However, there is little variability in the variable and thus can't be picked up and included.

As seen below, the variable has only 2 observations for *Have Insurance claim against crime or theft*.

```
#Dropping the variable.
table(finscope_OutClean$claim_insfor_theft)

##
## Don't have insurance claim against theft/crim
##                                         5674
##   Have Insurance claim against crime or theft
##                                            2

finscope_OutClean <- subset(finscope_OutClean, select = -claim_insfor_theft)
```

When we run our model below, which will act as the basis model we build on top of, the model runs fine and we don't run into any issues.

```
model1 <- lm(bus_net_profmnthly~.,data= finscope_OutClean)
summary(model1)

##
## Call:
## lm(formula = bus_net_profmnthly ~ ., data = finscope_OutClean)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -757579   -4862     425    5828 1183432
##
## Coefficients: (1 not defined because of singularities)
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 50730 on 2552 degrees of freedom
##   (3046 observations deleted due to missingness)
## Multiple R-squared:  0.5416, Adjusted R-squared:  0.5277
## F-statistic: 39.15 on 77 and 2552 DF,  p-value: < 2.2e-16
```

OUTLIER DETECTION AND REMOVAL IF APPROPRIATE:

Definition of outliers

Now that we have decided on a base model to make use of to understand the important predictors of business performance. We will now shift our attention to outlier detection using this model as the basis. By outliers, we are specifically referring to those data points in the data we are building our model with that tend to deviate significantly from the rest of the data points. The presence of such points can give us issues in model building as they tend to sway the model fitted into the data towards themselves, and thus giving us a false sense in trying to figure out which variables are important. If the outliers are not properly handled, they can skew the results of our model leading to invalid hypothesis tests that were run on the statistics estimated from the data.

There are many ways to think about outliers and the effects they can have on our model. For one, we can have the so-called global outliers. These sorts of outliers are typically due to some measurement errors, data entry errors, or some very unusual events that deviate significantly from other values. Usually, we usually just remove these from our model and replace them with some imputed value or we let go completely of that row by doing lit-wise deletion as they tend to be obvious errata or very rare events which give little insights about the general outlook of things.

Another view of outliers relates to the so-called contextual outliers, and it is highly likely that contextual outliers will be a high feature in this analysis. These outliers, differing from the ones we have outlined above are not due to errors or extremely rare events but rather have to do with the dataset itself. These variables are only outliers in a specific context but overall their behavior is normal or can be explained. For instance, when we look at this data set which consists mostly of own account enterprises, when the enterprise in question is a medium enterprise, they might exhibit high deviance in their values, and profits, but are overall normal when we consider them in the context of medium enterprises since medium enterprises generally rake in high profits. However, when these medium enterprises are grouped with their own account enterprises, they tend to exhibit characteristics of outliers.

Now, there are many operations we will run in this analysis to establish if certain data points are outliers or not. Once we have established that there are outliers, we have to further establish if those are influential outliers or not. It is usually the case that though certain points are outliers, however, they are not influential or significantly sway the model. Then, if that is the case, then they are not considered an issue. But what are influential points? Influential points are the points that have a large impact on the regression results. They have the potential effect of swaying the slope, significance of the variable, and the regression line fit if they are removed or added to the model.

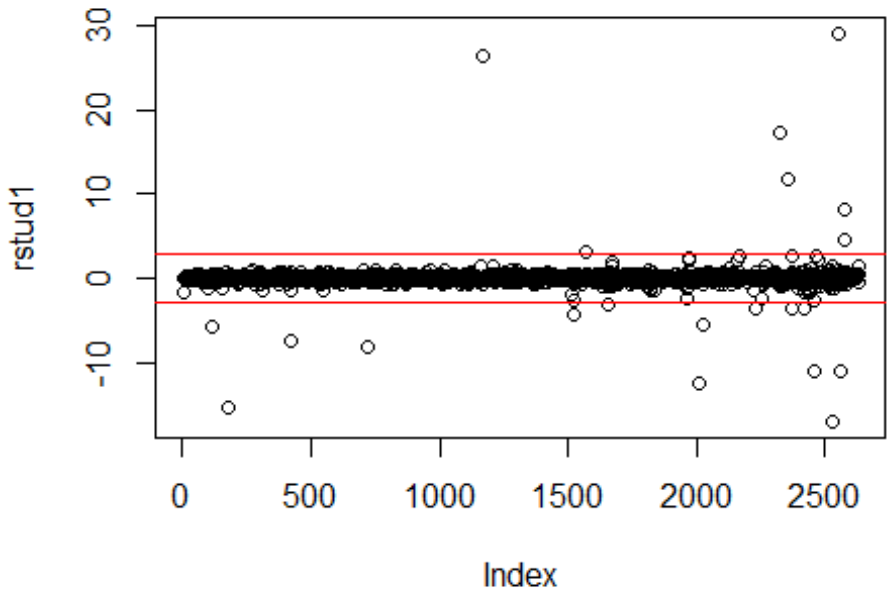
Note, that the are generally two ways in which values can be outliers, they can either be outliers in the explanatory variable or the y-variable, i.e. have high business profits. Or outliers in the x-variables/predictor variables which are called high leverage points, which are usually problematic and can have more influence over the regression fit line. If you have a mixture of outliers that are deviant in the y-axis and are also high leverage points, then we have a potentially dangerous concoction, which has the potential to be influential and, thus can sway the model significantly.

Studentised residuals

For our analysis, let's first look at outliers in general, without checking if they are influential or not. To do this, we make use of a method called the studentized residuals/t-distribution which works by dividing the residuals (the residual is simply the difference between the observations and its predicted value from the model) by the estimated std deviation of the sampling distribution of the residuals, where if the residuals are significantly larger than the std dev, tells us that the data point is an outlier. We generally expect the ratio of the residual to the std dev to be relatively a small value, and if we observe a value larger than absolute 3, more succinctly, a value that is more than three std deviations away from the mean of the sampling distributions of the residuals, then we have an outlier.

The code below works to detect such outliers from our model and appropriately deal with them.

```
#Get the t-values
rstud1 <- rstudent(model1)
plot(rstud1)

#Put an abline to signify the significance points.
abline(a=3, b=0, col="red")
abline(a=-3,b=0, col= "red")

identify(rstud1)
```




```
## integer(0)
```

Using identity () function and the which() function, we are able to identify the values that are potential outliers, and we can see this is an extensive list and its highly likely that they are contextual outliers. Thus, to ascertain this, we will do a visual inspection of these values to see if they are not outliers as a result of errors in the data collection process.

```
#Get all the values who t-values are above 3 in absolute terms
which(abs(rstud1) > 3)

##  205  312  797 1409 2339 3047 3136 3321 4166 4194 4667 4877 4977 4999 5113
5236
##  117  177  423  718 1167 1519 1570 1659 2014 2027 2233 2324 2354 2370 2421
2458
## 5412 5471 5503 5565 5570
## 2531 2550 2560 2577 2580

finscope_Outliers <- finscope_OutClean[c(205, 312, 797, 1409, 2339, 3047,
3136, 3321, 4166, 4194, 4667, 4877, 4977, 4999, 5113, 5236, 5412, 5471, 5503,
5565, 5570,117, 177, 423, 718, 1167, 1519, 1570, 1659, 2014, 2027, 2233,
2324, 2354, 2370, 2421, 2458, 2531, 2550, 2560, 2577, 2580),]

view(finscope_Outliers)
```

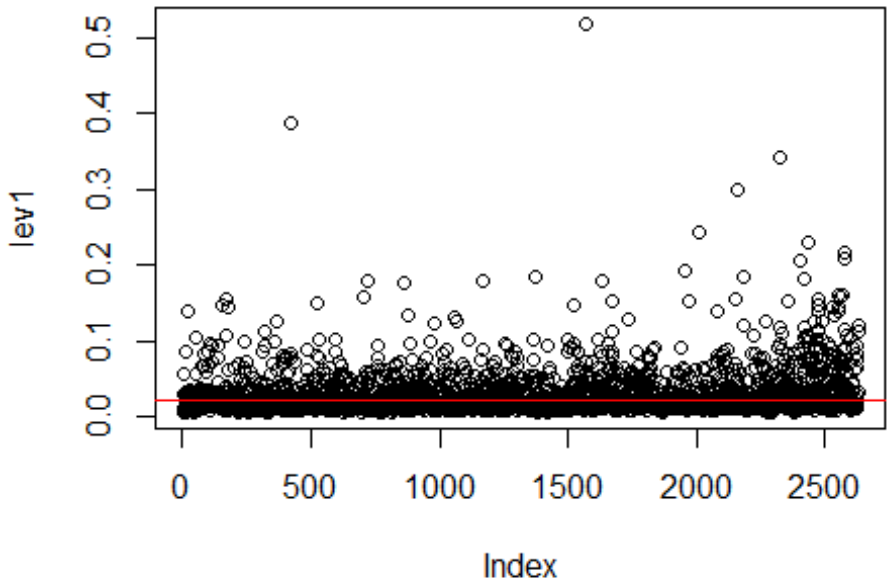
Upon visual inspection, indeed we don't see any obvious errors from the data, from data collection or recording, thus there are no global outliers. However, more diagnostics need

to be done to conclude that these values are influential and thus unduly pull our model, leading to skewing of the hypothesis tests run on the parameters estimated from this data.

High Leverage Points

First, we check for high leverage points, to see if there is no deviant predictor values. High leverage points have the potential of swaying the line of best fit and thus affect the hypothesis tests on the parameters.

```
#However, more diagnostics need to be done to conclude that these values are
influential and
#Unduly pulls our model.


# Let's check for high leverage values:
lev1 <- hatvalues(model1)
thresh <-  4*(29/5676)

plot(lev1)
abline(h=thresh, col = "red")
```



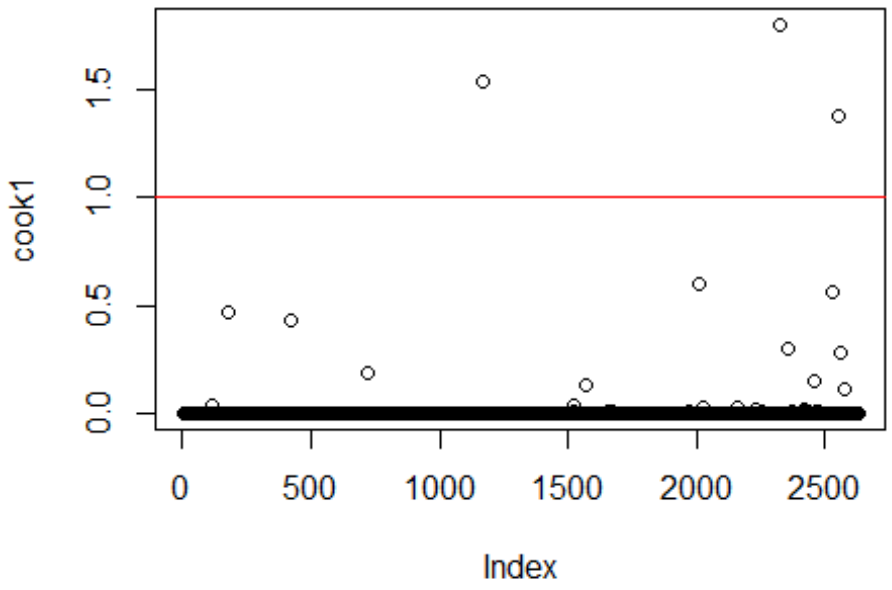
From the leverage plot above, it doesn't really give any sense of which points are high leverage points, and it is very unstable as it gives many points as being high leverage. The abline is also not giving any insights, even after relaxing the constraints. Thus, more still needs to be established beyond looking at leverage.

Cooks Distance

Thus, next we look at influence by looking at the cook's distance. Cook's distance is a measure of influence, which is a composite measure of outlierness on the y-axis and high leverage. Points that are influential have a high possibility of swaying the model thus affecting the hypothesis tests. As a rule of thumb, points that have a cook's distance that is bigger than 1 are deemed to be sufficiently influential such that they should be removed from the model.

The code below calculates the cook's distance, and points above the red abline will be deemed as influential, and thus upon inspection and testing, they will be removed from the model.

```
cook1 <- cooks.distance(model1)
plot(cook1)
abline(a= 1, b=0,col = "red")
identify(cook1)
```




```
## integer(0)

which(cook1 > 1)

## 2339 4877 5471
## 1167 2324 2550
```

Outlier Removal

Now, we have our influential points using cook distance. We see three values that are above the 1 threshold for the cook's distance and thus will be removed from the model. Below, we run a visual inspection of these values to assess what they are and if there's any context we can derive from looking at them before we move further in our analysis.

```
finscope_outliers <- finscope_OutClean[c(1167,2324,2550),]
head(finscope_outliers)
```

```
## # A tibble: 3 × 35
##   location     province   businesstype     wherebusoperate ownrent
tot_hours_wrk
##   <fct>        <fct>      <fct>            <fct>           <fct>
<dbl>
## 1 Urban formal Mpumalanga Render other se… Farm/small hol… Own th…
9
## 2 Rural formal Gauteng    Sell something … Residential pr… Don't …
6
## 3 Urban formal Gauteng    Render other se… Residential pr… Don't …
8
## # ℹ 29 more variables: tot_num_workers <dbl>, priv_indv <fct>,
## #   other_small_bus <fct>, other_large_bus <fct>, gov <fct>, tender_succ
<fct>,
## #   bank_loan <fct>, bus_turn_monthly <dbl>, bus_net_profmnthly <dbl>,
## #   age <dbl>, highlevel_edu <ord>, is_bus_only_src_inc <fct>,
## #   keep_fin_rec <fct>, ovrall_fin_access <fct>,
## #   having_security_measures <fct>, suffered_crimeortheft <fct>,
## #   is_bus_registered <fct>, submitted_tender_prop <fct>, …
```

```
view(finscope_outliers)
```

There's really no major difference between these values and other values in the dataset that make them seem to have an undue influence in the model beyond the fact that they are being singled out by cook's distance.

Here below we remove these outliers values from our model and thus conclude our work on dealing with outliers. And next, we deal with multicollinearity in the data.

```
finscope_Outclean2<- finscope_OutClean[-c(1167,2324,2550),]
```

CORRELATION ANALYSIS:

Next, we will run some correlation analysis to figure out which of the variables are correlated significantly. Correlation shows the extent to which two variables are associated with each other, that is, when variables change, the other variables change in response to the correlated variable. Note, that the could either be a negative correlation where an increase in one variable leads to a decline in another, or a positive correlation where an increase in one variable is associated with an increase in another. We use the scale of -1 to 1 to explain correlations, with values close to 1 or -1 associated with a strong

correlation/association with lesser values in absolute terms (that is closer to 0 in absolute terms) associated with lower intensity in correlations. Note, one thing about correlations, they only tell us about the above-mentioned relation, but do not give us any conclusive causal connection between the two variables, hence it then becomes important to establish the true nature of the relationship between the two variables, and not rely on the correlation value as many issues can be concealed by a high correlation value. In essence, there's a famous phrase, *correlation does not imply causation*.

One such issue that comes with high correlations and is detrimental to the proper function of our model, is the issue of multicollinearity. This occurs in a regression model when one of the predictor variables is highly correlated or is highly linearly related to one another in the model. Though correlation between predictor variables is expected, multicollinearity refers to an excessive form of this correlation which could be a result of one of these variables are simple linear combination of each other. If there is an instance of multicollinearity in the model, though the overall predictive power of the model will be intact, there could be an issue with individual variables and their stability in the model. Multicollinearity in the predictor variables can imply that one of the variables is redundant as the presence of one other variable has done all the explanatory work of the two variables. The presence of this multicollinearity can affect the choice of predictors in the final model, cloud our judgment on the precise effect of certain variables on our predicted variable, and can cause the coefficient of predictors to be unstable in the presence of multicollineraity thus less reliable, more sensitive and erratic to minor changes in the model, i.e the removal of variables.

To detect collinearity and by extension multicollinearity, we will be making use of the Varianace Inflation Factor (VIF) technique. The VIF is the ratio of 1 over the difference between one and the coefficient of determination (R^2). Note, that the coefficient of determination measures the amount of variation in the model that can be explained by the model relative to the simplest model possible, which is the mean model. If R is close to 1, then that model explains a lot of the variation in relation to the base model. Now, VIF works this way, it alternately makes each of the predictor variables the y-variable and calculates R^2 for each of these models. Then if the ratio is larger than 10, then this calls for concern, or in other words it indicates a high correlation. Why?, well the more difference between 1 and R^2, the bigger the value of R^2 indicating that one of the variables in the model is highly correlated with the explained variable, thus the more the difference, the bigger the value it'll be divided by 1. Thus, the bigger the value of VIF, the bigger the correlation between the x-variable in that particular iteration and the other x-variable in the predictor space.

```
model2 <- lm(bus_net_profmnthly~.,data= finscope_Outclean2)

#vif(model2)
```

From the code above, we are getting an error that disallows us from using the function to find highly correlated variables. Turns, the error is due to having a variable that has a R^2 of 1, which is a perfect correlation, thus crashing the equation. This is due to having a variable which is a linear combination of another variable in the model, thus leading to a perfect correlation. VIF is not powerful enough to handle perfect correlation, though it can handle

high correlation. To sort this issue, we use the *alias* function to find such variables and possibly remove them.

```
alias(model2)
```

```
## Model :
## bus_net_profmnthly ~ location + province + businesstype + wherebusoperate
+
##     ownrent + tot_hours_wrk + tot_num_workers + priv_indv +
other_small_bus +
##     other_large_bus + gov + tender_succ + bank_loan + bus_turn_monthly +
##     age + highlevel_edu + is_bus_only_src_inc + keep_fin_rec +
##     ovrall_fin_access + having_security_measures + suffered_crimeortheft +
##     is_bus_registered + submitted_tender_prop + comp_fin_recs +
##     owner_rentor_own_priv_res + expto_cust + supp_outof_sa +
##     have_ins + off_good_cred + have_accesstohowmanybusfuncs +
##     enterprise_classification + have_acces_to_howmany_ins_prod +
##     source_ofcredit + age_of_business
##
## Complete :
##
(Intercept)
## submitted_tender_propSubmitted a tender application last 12 months 0
##
locationTribal area
## submitted_tender_propSubmitted a tender application last 12 months 0
##
locationUrban formal
## submitted_tender_propSubmitted a tender application last 12 months 0
##
locationUrban informal
## submitted_tender_propSubmitted a tender application last 12 months 0
##
provinceFree State
## submitted_tender_propSubmitted a tender application last 12 months 0
##
provinceGauteng
## submitted_tender_propSubmitted a tender application last 12 months 0
##
provinceKZN
## submitted_tender_propSubmitted a tender application last 12 months 0
##
provinceLimpopo
## submitted_tender_propSubmitted a tender application last 12 months 0
##
provinceMpumalanga
## submitted_tender_propSubmitted a tender application last 12 months 0
##
provinceN.Cape
## submitted_tender_propSubmitted a tender application last 12 months 0
```

```
##
provinceN.West
## submitted_tender_propSubmitted a tender application last 12 months 0
##
provinceW.Cape
## submitted_tender_propSubmitted a tender application last 12 months 0
##
businesstypeRear livestock/poultry and sell e.g. chickens
## submitted_tender_propSubmitted a tender application last 12 months 0
##
businesstypeRender a professional service e.g. doctor, lawyer, accountant,
engineer, consultant
## submitted_tender_propSubmitted a tender application last 12 months 0
##
businesstypeRender a skilled service e.g. mechanic, plumber, hair salon,
barber, painting, landscaping
## submitted_tender_propSubmitted a tender application last 12 months 0
##
businesstypeRender building/construction services
## submitted_tender_propSubmitted a tender application last 12 months 0
##
businesstypeRender other services e.g. car wash, garden services, transport
(taxi services), catering
## submitted_tender_propSubmitted a tender application last 12 months 0
##
businesstypeRender tourism-related services e.g.
accommodation/hotel/B&B/guest house, tour operators
## submitted_tender_propSubmitted a tender application last 12 months 0
##
businesstypeSell by-products of animals e.g. meat, eggs, milk
## submitted_tender_propSubmitted a tender application last 12 months 0
##
businesstypeSell something in the same form that I buy from someone else
(dont add value, e.g. cigarettes)
## submitted_tender_propSubmitted a tender application last 12 months 0
##
businesstypeSell something that I buy but add value to, e.g. repackage, cook,
etc
## submitted_tender_propSubmitted a tender application last 12 months 0
##
businesstypeSell something that I collect from nature, e.g. herbs, firewood,
charcoal, thatch, sand, stone
## submitted_tender_propSubmitted a tender application last 12 months 0
##
businesstypeSell something that I get for free, e.g. second hand clothes,
scrap metal
## submitted_tender_propSubmitted a tender application last 12 months 0
##
businesstypeSell something that I make e.g. crafts, clothes, furniture,
bricks
```

```
## submitted_tender_propSubmitted a tender application last 12 months 0
##
wherebusoperateCar/truck/vehicle
## submitted_tender_propSubmitted a tender application last 12 months 0
##
wherebusoperateDoor to door/Go to customers
## submitted_tender_propSubmitted a tender application last 12 months 0
##
wherebusoperateFarm/small holding
## submitted_tender_propSubmitted a tender application last 12 months 0
##
wherebusoperateOffice block/office park
## submitted_tender_propSubmitted a tender application last 12 months 0
##
wherebusoperateOnline - internet, phone selling
## submitted_tender_propSubmitted a tender application last 12 months 0
##
wherebusoperateResidential premises - dwelling/garage/building on residential
premises
## submitted_tender_propSubmitted a tender application last 12 months 0
##
wherebusoperateSchool Cafeteria
## submitted_tender_propSubmitted a tender application last 12 months 0
##
wherebusoperateShopping mall
## submitted_tender_propSubmitted a tender application last 12 months 0
##
wherebusoperateStall/table/container in a designated trading or market area
## submitted_tender_propSubmitted a tender application last 12 months 0
##
wherebusoperateStreet/street corner/pavement
## submitted_tender_propSubmitted a tender application last 12 months 0
##
ownrentOwn the business premises
## submitted_tender_propSubmitted a tender application last 12 months 0
##
tot_hours_wrk
## submitted_tender_propSubmitted a tender application last 12 months 0
##
tot_num_workers
## submitted_tender_propSubmitted a tender application last 12 months 0
##
priv_indvDoesn't sell to Private Individuals
## submitted_tender_propSubmitted a tender application last 12 months 0
##
other_small_busSells to other small enterpises
## submitted_tender_propSubmitted a tender application last 12 months 0
##
other_large_busSells to larger enterpises
## submitted_tender_propSubmitted a tender application last 12 months 0
```

```
##
govSells to government
## submitted_tender_propSubmitted a tender application last 12 months 0
##
tender_succNo
## submitted_tender_propSubmitted a tender application last 12 months 1
##
tender_succYes
## submitted_tender_propSubmitted a tender application last 12 months 1
##
bank_loanYES
## submitted_tender_propSubmitted a tender application last 12 months 0
##
bus_turn_monthly
## submitted_tender_propSubmitted a tender application last 12 months 0
##                                                                   age
## submitted_tender_propSubmitted a tender application last 12 months 0
##
highlevel_edu.L
## submitted_tender_propSubmitted a tender application last 12 months 0
##
highlevel_edu.Q
## submitted_tender_propSubmitted a tender application last 12 months 0
##
highlevel_edu.C
## submitted_tender_propSubmitted a tender application last 12 months 0
##
highlevel_edu^4
## submitted_tender_propSubmitted a tender application last 12 months 0
##
highlevel_edu^5
## submitted_tender_propSubmitted a tender application last 12 months 0
##
highlevel_edu^6
## submitted_tender_propSubmitted a tender application last 12 months 0
##
highlevel_edu^7
## submitted_tender_propSubmitted a tender application last 12 months 0
##
is_bus_only_src_incBusiness is the only source of income
## submitted_tender_propSubmitted a tender application last 12 months 0
##
keep_fin_recKeep financial records
## submitted_tender_propSubmitted a tender application last 12 months 0
##
ovrall_fin_accessFormal
## submitted_tender_propSubmitted a tender application last 12 months 0
##
ovrall_fin_accessInformal
## submitted_tender_propSubmitted a tender application last 12 months 0
```

```
##
ovrall_fin_accessNot served
## submitted_tender_propSubmitted a tender application last 12 months 0
##
having_security_measuresHave security Measures
## submitted_tender_propSubmitted a tender application last 12 months 0
##
suffered_crimeortheftBusiness suffered crime or theft in the last 12 months
## submitted_tender_propSubmitted a tender application last 12 months 0
##
is_bus_registeredBusiness is registered
## submitted_tender_propSubmitted a tender application last 12 months 0
##
comp_fin_recsBusiness keeps computerized financial records
## submitted_tender_propSubmitted a tender application last 12 months 0
##
owner_rentor_own_priv_resOwn the private residence
## submitted_tender_propSubmitted a tender application last 12 months 0
##
expto_custExport to outside of SA
## submitted_tender_propSubmitted a tender application last 12 months 0
##
supp_outof_saHave suppliers out of SA
## submitted_tender_propSubmitted a tender application last 12 months 0
##
have_insHave Insurance
## submitted_tender_propSubmitted a tender application last 12 months 0
##
off_good_credYes always
## submitted_tender_propSubmitted a tender application last 12 months 0
##
off_good_credYes, sometimes
## submitted_tender_propSubmitted a tender application last 12 months 0
##
have_accesstohowmanybusfuncs
## submitted_tender_propSubmitted a tender application last 12 months 0
##
enterprise_classification.L
## submitted_tender_propSubmitted a tender application last 12 months 0
##
enterprise_classification.Q
## submitted_tender_propSubmitted a tender application last 12 months 0
##
enterprise_classification.C
## submitted_tender_propSubmitted a tender application last 12 months 0
##
have_acces_to_howmany_ins_prod
## submitted_tender_propSubmitted a tender application last 12 months 0
##
source_ofcreditBorrowing from friends/family
```

```
## submitted_tender_propSubmitted a tender application last 12 months 0
##
source_ofcreditFormal
## submitted_tender_propSubmitted a tender application last 12 months 0
##
source_ofcreditInformal
## submitted_tender_propSubmitted a tender application last 12 months 0
##
source_ofcreditNot served
## submitted_tender_propSubmitted a tender application last 12 months 0
##
age_of_business
## submitted_tender_propSubmitted a tender application last 12 months 0
```

From the code above, it seems having submitted a tender application in the last 12 months and having a tender application being yes or no are perfectly correlated with each other. This makes sense, since in order for your tender application to either be successful or not, it must have been submitted anyway, which renders these two variables perfectly correlated with each other.Also, since the variable *tender_succ* also makes provision for the instead of not having submitted a tender proposal, this renders the *submitted_tender_prop* variable useless since the above variable makes all the provisions. See the code below where the categories on *Did not submit a tender application in the last 12 months* and *Did not apply* are equal.

```
table(finscope_Outclean2$tender_succ)

##
## Did not apply            No           Yes
##          5476           127            70

table(finscope_OutClean$submitted_tender_prop)

##
## Did not submit a tender application in the last 12 months
##                                                      5479
##             Submitted a tender application last 12 months
##                                                       197
```

Thus, from our model we will remove the variable on having *submitted a tender application*, as we feel it should be made redundant by the other variable, and we believe that companies that being successful or not is a much more robust predictor of business performance than mere applying. And since the remaining has also taken into consideration the categories of the original variable we are removing, then we have dealt with the redundancy.

```
finscope_Outclean2 <- subset(finscope_OutClean, select = -
submitted_tender_prop)
```

Now that we have removed the problematic variable from the model, we can now run the model and detect any redundancy or highly correlated variables in the model which can cause some of the variables to be unstable thus affect the significance tests.

```
model3 <-  lm(bus_net_profmnthly~.,data= finscope_Outclean2)
summary(model3)
```

```
##
## Call:
## lm(formula = bus_net_profmnthly ~ ., data = finscope_Outclean2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -757579   -4862     425    5828 1183432
##



## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 50730 on 2552 degrees of freedom
##   (3046 observations deleted due to missingness)
## Multiple R-squared:  0.5416, Adjusted R-squared:  0.5277
## F-statistic: 39.15 on 77 and 2552 DF,  p-value: < 2.2e-16
```

```
vif(model3)
```

```
##                                     GVIF Df GVIF^(1/(2*Df))
## location                        1.907223  3        1.113611
## province                        2.306768  8        1.053629
## businesstype                    3.800829 12        1.057211
## wherebusoperate                 4.773419 10        1.081288
## ownrent                         2.637809  1        1.624133
## tot_hours_wrk                   1.233785  1        1.110759
## tot_num_workers                 5.334603  1        2.309676
## priv_indv                       1.287998  1        1.134900
## other_small_bus                 1.436707  1        1.198627
## other_large_bus                 1.681318  1        1.296657
## gov                             1.313056  1        1.145886
## tender_succ                     1.495840  2        1.105914
## bank_loan                       1.344053  1        1.159333
## bus_turn_monthly                1.317095  1        1.147648
## age                             1.608871  1        1.268413
## highlevel_edu                   2.870896  7        1.078240
## is_bus_only_src_inc             1.131883  1        1.063900
## keep_fin_rec                    1.582877  1        1.258124
## ovrall_fin_access               2.333717  3        1.151705
```

```
## having_security_measures       1.900320  1        1.378521
## suffered_crimeortheft          1.083344  1        1.040838
## is_bus_registered              1.853504  1        1.361434
## comp_fin_recs                  2.250682  1        1.500227
## owner_rentor_own_priv_res      1.721283  1        1.311977
## expto_cust                     1.169347  1        1.081364
## supp_outof_sa                  1.183391  1        1.087838
## have_ins                       2.367815  1        1.538771
## off_good_cred                  1.331825  2        1.074266
## have_accesstohowmanybusfuncs   2.079654  1        1.442100
## enterprise_classification      9.101826  3        1.444956
## have_acces_to_howmany_ins_prod 1.882253  1        1.371952
## source_ofcredit                2.682721  4        1.131285
## age_of_business                1.298622  1        1.139571
```

From the model we have run above, we see that two variables have concerning VIF values which are above 5, which are *enterprise_classification* and *tot_num_workers*. Now it makes logical sense why these two variables are problematic in our analysis. This pertains to the fact that *enterprise_classification* was constructed from the *tot_num_workers* variable, hence it makes sense that we have such a high VIF score. We will remove the *tot_num_workers* variable since we are interested in analyzing the different classes of enterprises and the effect each has on business performance. This would be highly valuable to policymakers and or investors who are interested in the size of businesses to invest in and the support required by each.

```
finscope_Outclean3 <- subset(finscope_Outclean2, select = -tot_num_workers)
```

Nice, now we have dealt with perfectly correlated values and other variables that have high multicollinearity and thus introduce redundancy, we can then move forward with our analysis.

Dealing with missing values using imputation…

What is Imputation

Now we shift our focus to dealing with the missing values in our dataset and ways to solve this issue. There are many reasons why data may be missing, one may arise due to mistakes in the data collection process, data privacy concerns, or the most common, which is item non-response where a participant refuses to respond. Now the reason for the non-response is also as important, but essentially, we hope that there is no pattern in the reasons why the respondents chose to omit the question, thus the data is missing completely at random.

There are many ways to deal with missing data, assuming that the data is missing at random. One of the ways is to apply litwise deletion which deletes all the cases of missing observations, though some rows may have missing data in only a few columns, litwise deletion will remove the entire row. Thus, it is clear utilizing litwise deletion can throw away a lot of valuable information which decreases the statistical power of our analysis. Further, if the data is not missing randomly, deletion of a portion of the data may weaken our model's ability to predict/model certain instances of the world, also lead to biases in

the coefficients of our model and thus a misleading view of the world we are trying to understand.

Thus, this supports the logic for imputation is that some columns may be missing data in a few rows and not other rows thus deleting some of these rows may lead to less data and more biases.

Imputation Process including the Predictive Mean Matching Procedure

There are many imputation procedures that can be used but this analysis will make use of Predictive Mean Matching imputation using the mice package in R. This method is intuitive, solves the issue of bias and avoids throwing away data which is useful in our analysis. Predictive Mean Matching is a very smart way of doing imputation. It does this by running a regression model on the column with missing values as the target variable, using the other columns to run this regression with as the explainers. If these columns have missing values too, the means of those columns are used as the placeholder whilst regressing over the other column that is the target in that instance, then those means are replaced when its that column's turn is predicted. For the random element in the data, it is assumed that the errors are random, independent, and identically distributed and their distribution follows a a normal distribution, and past values are used to pick the random element. This process is run multiple times and is shown as *maxit = 60* in the mice formula, where the iterative is done 60 times to have robust and unbiased regression errors.

Now the process I have specified above is the stochastic regression for imputing values and it has some serious drawbacks, one, the model can predict some implausible values, such as negative values. Also, if the underlying data is heteroscedastic, the model above would fail to pick it up as it would assume that the stochastic element variance is constant over time when it is not. i.e. it will have given x, assume the variance remains constant over x, whereas in reality, it may be that as x changes, the variance changes.

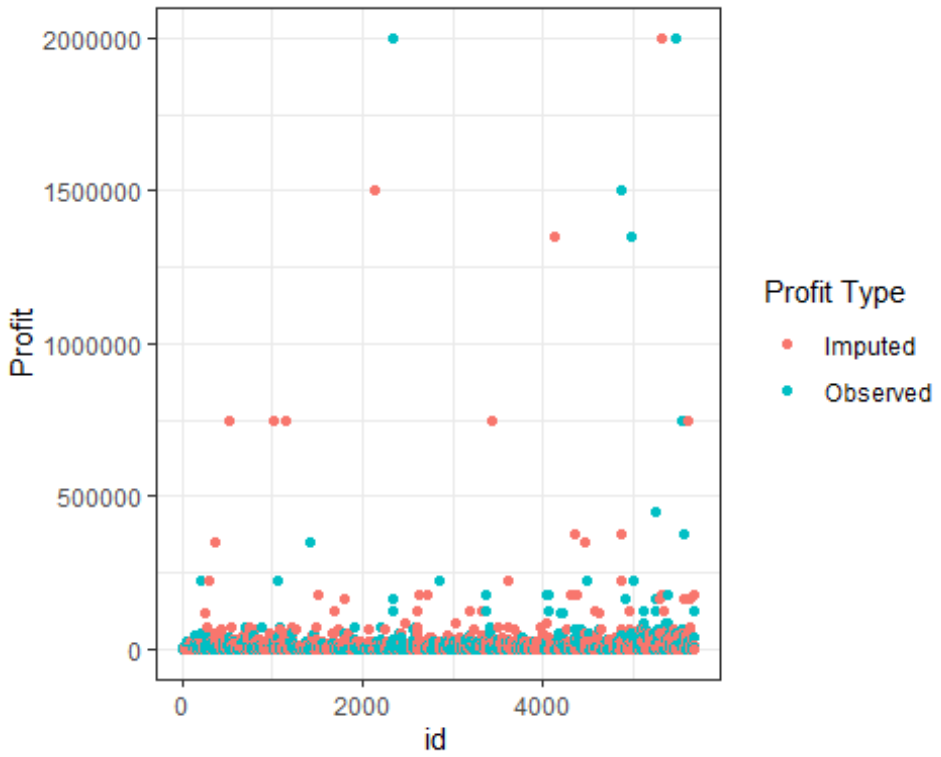
PMM is a major innovation over the stochastic regression imputation as it after doing the predictions, then on the basis of the predictions, will look for similar predicted values, group them, and then randomly draw from them to replace the missing value. This ensures that the original distribution of the data is preserved.

Note, the process is the same for categorical data, and one major difference pertains to the model used to predict the missing values, with numerical data using the regression model whilst the categorical data using the logistic data if it is a binary and the multi-modal logistic regression model for the multi-category instance.

```
imputed_finscope_data <- mice(finscope_Outclean3,m= 1,maxit = 60,printFlag =
FALSE, seed = 1234)

## Warning: Number of logged events: 60

imputed_finscope_data = complete(imputed_finscope_data, "long")
```

Now, lets visualize the distribution of the original data and the imputed values to show that indeed that the data has preserved its distribution.

```r
# Add a row number column to both datasets
finscope_Outclean3 <- finscope_Outclean3 %>% mutate(row = row_number())
imputed_finscope_data <- imputed_finscope_data %>% mutate(row = row_number())

# Join the original and imputed datasets by row number
joined_data <- left_join(finscope_Outclean3, imputed_finscope_data, by =
"row", suffix = c("_orig", "_imputed"))

# Create a new column to indicate whether the profit value is observed or
imputed
joined_data$profit_type <- ifelse(is.na(joined_data$bus_net_profmnthly_orig),
"Imputed", "Observed")

# Create a scatter plot of profit versus time, with different colors for
observed and imputed values
ggplot(joined_data, aes(x = .id, y = bus_net_profmnthly_imputed, color =
profit_type)) +
  geom_point() +
  labs(x = "id", y = "Profit", color = "Profit Type") +
  theme_bw()
```



Indeed from the plot above, we see that the two sets of data, the imputed and the observed value both show similar distribution and pattern, thus we consider our imputation exercise as a success..

Now that we are done with our imputation exercise, we shift to dropping some of the variables we created during the imputation process and we go on further with our analysis.

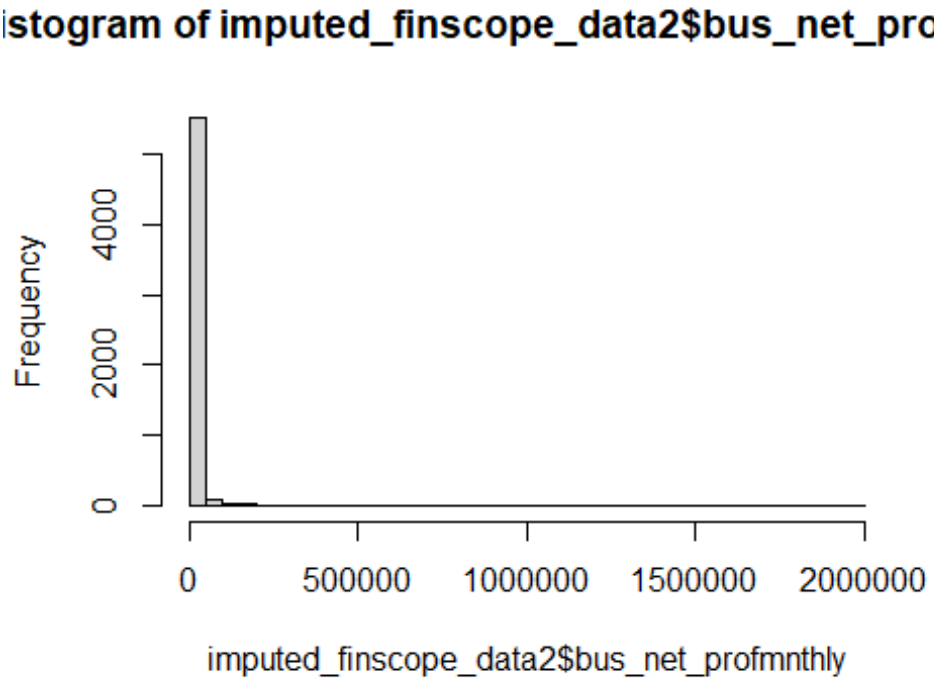
Fixing non-linearity..

Below, we will plot the numerical variables in our analysis to check for the normality assumptions on our variables by plotting their distributions. We will also check how far the largest values in the distributions are from the bulk of the data, which may lead to our distributions being asymmetrical, for example, larger values might pull the distribution towards the right, shield the distribution of the larger values. Thus, we might need to transform our variables or change the scale with which we are analyzing our data to get much more symmetric and normally distributed values.

To fix this issue, we might need to log-transform our variables of interest such that we dampen the effect of the large values of our variables in skewing our distributions. This transformation will then lead to a much more symmetric and normal distribution of the variable, which will then be more in line with the assumptions of our model that the variables are normally distributed. Furthermore, we will have a better view of our variable on the log-scale better than on the original scale. There is also the added bonus that the log transformation will have, which is, that it will tend to linearize any relationships that might be exponential or multiplicative. This transformation will also fix heteroscedasticity to some extent. Note, we don't use log in its original format, but add 1p, which is equivalent to adding 1 to each value. This is due to the fact that some of the values are equal to 0 of which the log of zero is not defined, and thus adding 1 at each instance will ensure that we don't have a log of zero.

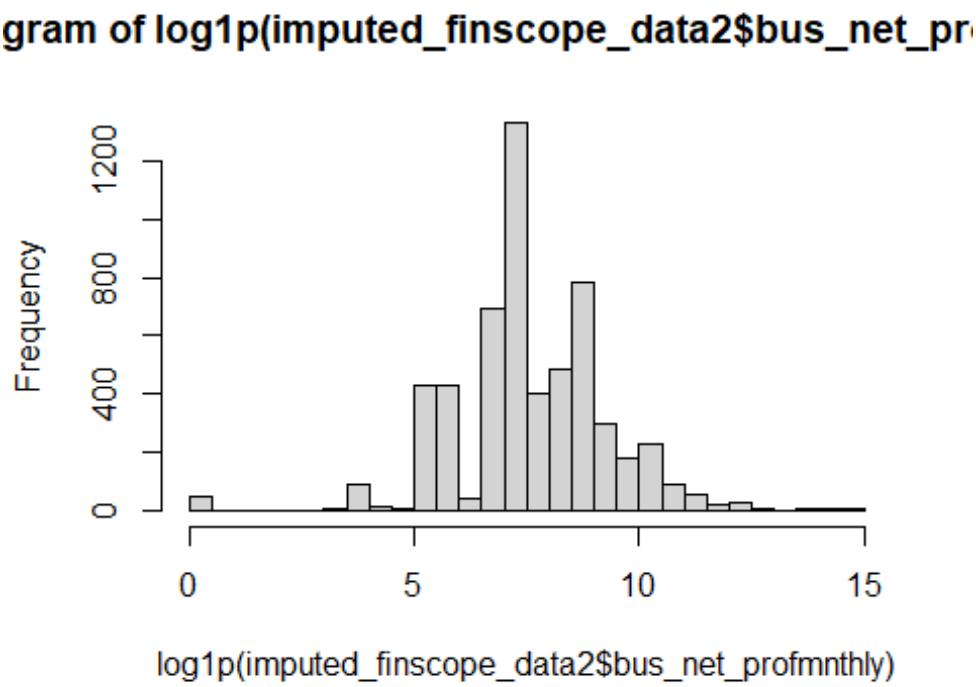
Below, we start with the business net profits monthly variable and check if they are normally distributed or not. From the distribution below, it is clear that the data is skewed to the right, i.e is pulled to the right by very large positive values. Also, we see that the largest values in our variable are very far from the bulk of the other values, thus we have high variability in our data, which causes the plot to not accurately depict the distribution of incomes. This might also be an issue with the other numerical variables.

```
hist(imputed_finscope_data2$bus_net_profmnthly,n=50)
```



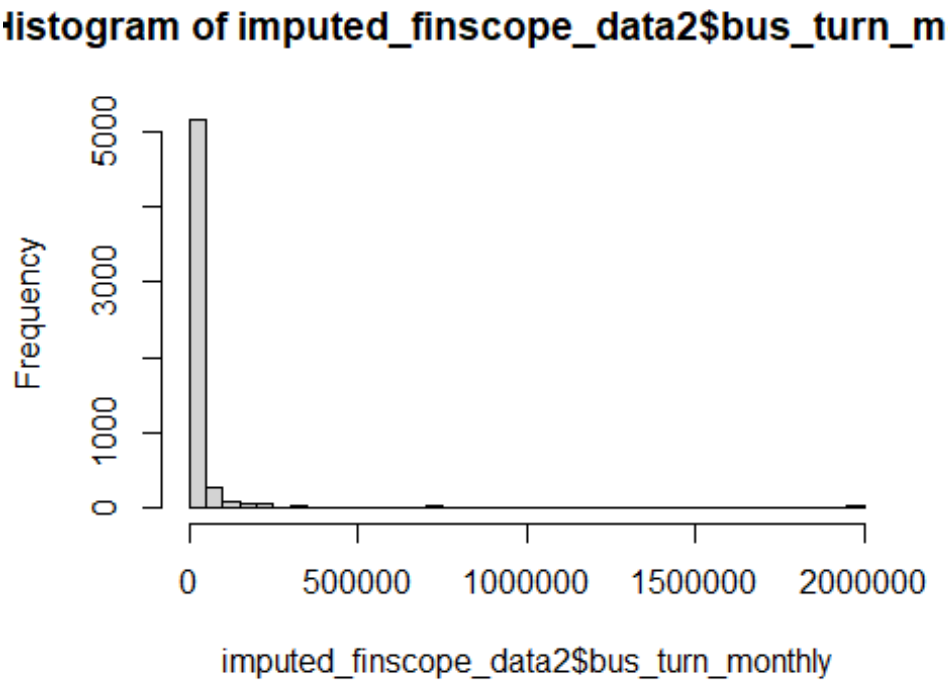
Below we have logged transformed our variable so that it's more in line with the model assumptions. From the plot below, we can see that indeed, the distribution is more reflective of a symmetrical and normal distribution, and the large values no longer skew the distribution towards being more skewed positively as much.

```
hist(log1p(imputed_finscope_data2$bus_net_profmnthly), n=50)
```



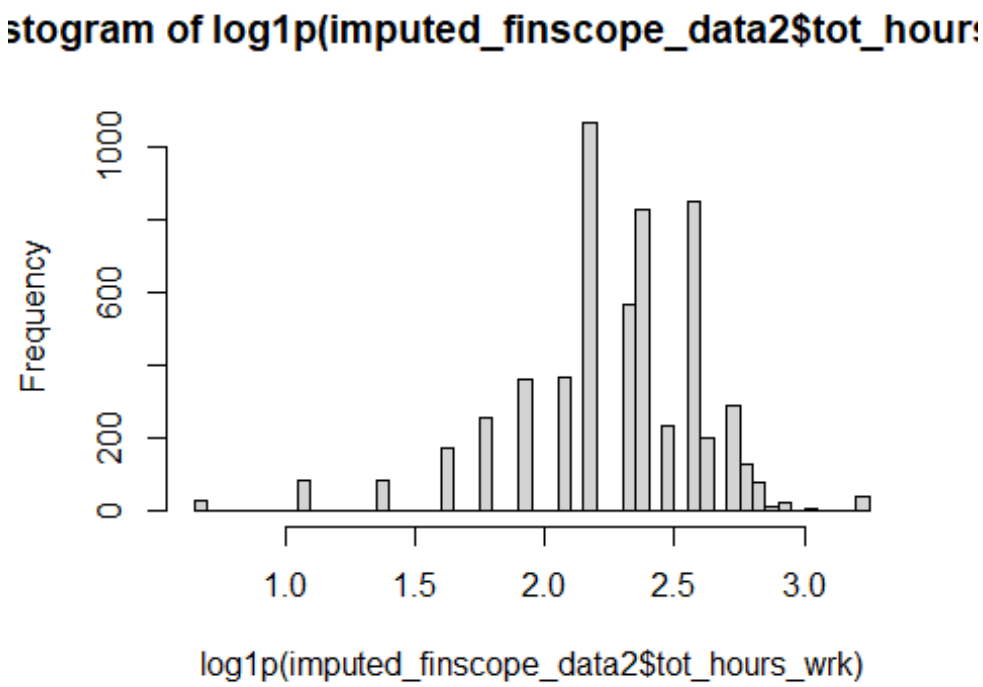
The same logic applies for businesses turnover per month, and as such the same transformations will be applicable and the effects are similar to the ones we ran above for profits.

```
hist(imputed_finscope_data2$bus_turn_monthly,n=50)
```



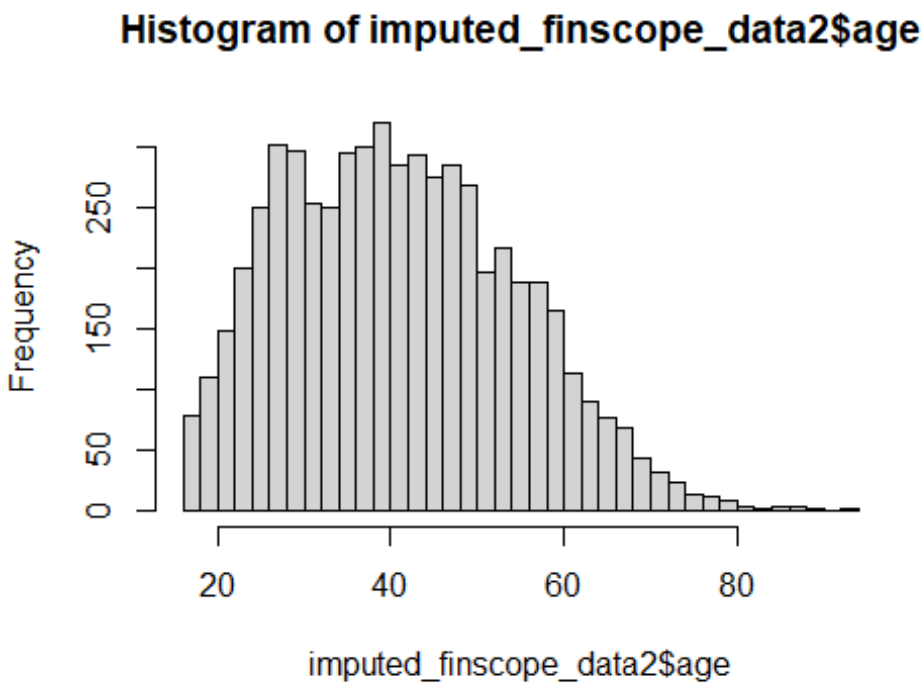
Also, this variable shows skewness and thus will also need to be log transformed to make it much more symmetrical and appropriate due to the statistical assumptions made by our model.

```
hist(log1p(imputed_finscope_data2$tot_hours_wrk),n=50)
```



This one also seems to be reasonably symmetrical and thus we will also not make any transformations on it.

```
hist(imputed_finscope_data2$age,n=50)
```



From the graphs we have plotted above, its clear which of the numerical variables show non-symmetrical shapes and thus will need to be logged transformed permanently from the data and thus be in line with our model assumptions.

We log transform two variables, namely, *business net profits monthly* and the *business net turnover monthly*.

```
imputed_finscope_data2$bus_net_profmnthly_log <-
log1p(imputed_finscope_data2$bus_net_profmnthly)
imputed_finscope_data2$bus_turn_monthly_log   <-
log1p(imputed_finscope_data2$bus_turn_monthly)
```

Here we drop the variables that were not transformed in favor of the log-transformed variables which are much more appropriate.

```
imputed_finscope_data4 <- subset(imputed_finscope_data2, select = -
c(bus_net_profmnthly,bus_turn_monthly))
```

Lets model once more

```
model.log <-  lm(bus_net_profmnthly_log~.,data= imputed_finscope_data4)

summary(model.log)

##
## Call:
```

```
## lm(formula = bus_net_profmnthly_log ~ ., data = imputed_finscope_data4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.2941  -0.6977   0.0763   0.7994   4.3579
##
## Coefficients:
##
Estimate
## (Intercept)
4.3865299
## locationTribal area
0.0807274
## locationUrban formal
0.1946157
## locationUrban informal
0.2059617
## provinceFree State
0.3794553
## provinceGauteng
0.2488712
## provinceKZN
0.3526534
## provinceLimpopo
0.1287899
## provinceMpumalanga
0.1612007
## provinceN.Cape
0.1673060
## provinceN.West
0.0381995
## provinceW.Cape
-0.0632253
## businesstypeRear livestock/poultry and sell e.g. chickens
-0.2680330
## businesstypeRender a professional service e.g. doctor, lawyer, accountant,
engineer, consultant            -0.2763161
## businesstypeRender a skilled service e.g. mechanic, plumber, hair salon,
barber, painting, landscaping     -0.2880394
## businesstypeRender building/construction services
0.3588993
## businesstypeRender other services e.g. car wash, garden services,
transport (taxi services), catering      -0.3643505
## businesstypeRender tourism-related services e.g.
accommodation/hotel/B&B/guest house, tour operators       -0.4645050
## businesstypeSell by-products of animals e.g. meat, eggs, milk
-0.1486149
## businesstypeSell something in the same form that I buy from someone else
(dont add value, e.g. cigarettes) -0.1849850
## businesstypeSell something that I buy but add value to, e.g. repackage,
```

```
cook, etc                            -0.3544745
## businesstypeSell something that I collect from nature, e.g. herbs,
firewood, charcoal, thatch, sand, stone -0.0727029
## businesstypeSell something that I get for free, e.g. second hand clothes,
scrap metal                        -0.2432904
## businesstypeSell something that I make e.g. crafts, clothes, furniture,
bricks                               -0.0442394
## wherebusoperateCar/truck/vehicle
0.1786614
## wherebusoperateDoor to door/Go to customers
-0.5022444
## wherebusoperateFarm/small holding
-0.3999632
## wherebusoperateOffice block/office park
-0.9579578
## wherebusoperateOnline - internet, phone selling
-1.4703762
## wherebusoperateOpen space-Isipingo
1.3600956
## wherebusoperateResidential premises - dwelling/garage/building on
residential premises                      -0.6118165
## wherebusoperateSchool Cafeteria
-0.5441550
## wherebusoperateShopping mall
-0.0838662
## wherebusoperateStall/table/container in a designated trading or market
area                                  -0.3894504
## wherebusoperateStreet/street corner/pavement
-0.6353497
## ownrentOwn the business premises
0.0189998
## tot_hours_wrk
0.0014299
## priv_indvDoesn't sell to Private Individuals
0.1230099
## other_small_busSells to other small enterpises
-0.0712921
## other_large_busSells to larger enterpises
0.1475466
## govSells to government
0.6262697
## tender_succNo
-0.0953256
## tender_succYes
0.3549231
## bank_loanYES
-0.2097883
## age
0.0008138
## highlevel_edu.L
```

```
0.5292684
## highlevel_edu.Q
-0.0624076
## highlevel_edu.C
-0.2795432
## highlevel_edu^4
-0.0891736
## highlevel_edu^5
0.5573479
## highlevel_edu^6
0.7376107
## highlevel_edu^7
0.4719196
## is_bus_only_src_incBusiness is the only source of income
0.0253505
## keep_fin_recKeep financial records
-0.0012173
## ovrall_fin_accessFormal
-0.1793188
## ovrall_fin_accessInformal
-0.2040601
## ovrall_fin_accessNot served
-0.1156189
## having_security_measuresHave security Measures
0.0868893
## suffered_crimeortheftBusiness suffered crime or theft in the last 12
months                                -0.1796772
## is_bus_registeredBusiness is registered
0.0245464
## comp_fin_recsBusiness keeps computerized financial records
0.0484706
## owner_rentor_own_priv_resOwn the private residence
-0.2180300
## expto_custExport to outside of SA
-0.0359154
## supp_outof_saHave suppliers out of SA
0.0565889
## have_insHave Insurance
0.0566013
## off_good_credYes always
-0.0463257
## off_good_credYes, sometimes
0.0436740
## have_accesstohowmanybusfuncs
-0.0440186
## enterprise_classification.L
0.4201720
## enterprise_classification.Q
-0.0324607
## enterprise_classification.C
```

```
-0.2835895
## have_acces_to_howmany_ins_prod
-0.0935527
## source_ofcreditBorrowing from friends/family
-0.2090315
## source_ofcreditFormal
0.3008441
## source_ofcreditInformal
-0.1243575
## source_ofcreditNot served
-0.1039837
## age_of_business
0.0056352
## bus_turn_monthly_log
0.4991417
##
Std. Error
## (Intercept)
0.2726077
## locationTribal area
0.0830343
## locationUrban formal
0.0754965
## locationUrban informal
0.0969934
## provinceFree State
0.0893064
## provinceGauteng
0.0723567
## provinceKZN
0.0726644
## provinceLimpopo
0.0845868
## provinceMpumalanga
0.0859890
## provinceN.Cape
0.1023417
## provinceN.West
0.0841700
## provinceW.Cape
0.0793910
## businesstypeRear livestock/poultry and sell e.g. chickens
0.1520582
## businesstypeRender a professional service e.g. doctor, lawyer, accountant,
engineer, consultant             0.1861450
## businesstypeRender a skilled service e.g. mechanic, plumber, hair salon,
barber, painting, landscaping      0.1086464
## businesstypeRender building/construction services
0.1832159
## businesstypeRender other services e.g. car wash, garden services,
```

```
transport (taxi services), catering       0.1211379
## businesstypeRender tourism-related services e.g.
accommodation/hotel/B&B/guest house, tour operators        0.1930273
## businesstypeSell by-products of animals e.g. meat, eggs, milk
0.1615380
## businesstypeSell something in the same form that I buy from someone else
(dont add value, e.g. cigarettes)  0.0993029
## businesstypeSell something that I buy but add value to, e.g. repackage,
cook, etc                           0.1103327
## businesstypeSell something that I collect from nature, e.g. herbs,
firewood, charcoal, thatch, sand, stone  0.1904503
## businesstypeSell something that I get for free, e.g. second hand clothes,
scrap metal                        0.2127382
## businesstypeSell something that I make e.g. crafts, clothes, furniture,
bricks                              0.1267535
## wherebusoperateCar/truck/vehicle
0.2685508
## wherebusoperateDoor to door/Go to customers
0.1451030
## wherebusoperateFarm/small holding
0.1781113
## wherebusoperateOffice block/office park
0.2104158
## wherebusoperateOnline - internet, phone selling
0.3963777
## wherebusoperateOpen space-Isipingo
1.4331841
## wherebusoperateResidential premises - dwelling/garage/building on
residential premises                       0.1105986
## wherebusoperateSchool Cafeteria
0.2553437
## wherebusoperateShopping mall
0.2061678
## wherebusoperateStall/table/container in a designated trading or market
area                                  0.1652721
## wherebusoperateStreet/street corner/pavement
0.1222124
## ownrentOwn the business premises
0.0617940
## tot_hours_wrk
0.0061790
## priv_indvDoesn't sell to Private Individuals
0.1433973
## other_small_busSells to other small enterpises
0.0679012
## other_large_busSells to larger enterpises
0.1152948
## govSells to government
0.1377220
## tender_succNo
```

```
0.1340717
## tender_succYes
0.1861236
## bank_loanYES
0.1377070
## age
0.0017717
## highlevel_edu.L
0.1208030
## highlevel_edu.Q
0.0974433
## highlevel_edu.C
0.1016808
## highlevel_edu^4
0.0729567
## highlevel_edu^5
0.0809664
## highlevel_edu^6
0.0947785
## highlevel_edu^7
0.0644663
## is_bus_only_src_incBusiness is the only source of income
0.0411929
## keep_fin_recKeep financial records
0.0468898
## ovrall_fin_accessFormal
0.1152915
## ovrall_fin_accessInformal
0.0894800
## ovrall_fin_accessNot served
0.0501961
## having_security_measuresHave security Measures
0.0703943
## suffered_crimeortheftBusiness suffered crime or theft in the last 12
months                                    0.0772959
## is_bus_registeredBusiness is registered
0.0635180
## comp_fin_recsBusiness keeps computerized financial records
0.0910815
## owner_rentor_own_priv_resOwn the private residence
0.0611132
## expto_custExport to outside of SA
0.2386964
## supp_outof_saHave suppliers out of SA
0.2635301
## have_insHave Insurance
0.0659926
## off_good_credYes always
0.0719795
## off_good_credYes, sometimes
```

```
0.0467027
## have_accesstohowmanybusfuncs
0.0255544
## enterprise_classification.L
0.3201338
## enterprise_classification.Q
0.2419883
## enterprise_classification.C
0.1462168
## have_acces_to_howmany_ins_prod
0.0391900
## source_ofcreditBorrowing from friends/family
0.1487361
## source_ofcreditFormal
0.3699511
## source_ofcreditInformal
0.2342417
## source_ofcreditNot served
0.0842194
## age_of_business
0.0028726
## bus_turn_monthly_log
0.0143350
##
t value
## (Intercept)
16.091
## locationTribal area
0.972
## locationUrban formal
2.578
## locationUrban informal
2.123
## provinceFree State
4.249
## provinceGauteng
3.440
## provinceKZN
4.853
## provinceLimpopo
1.523
## provinceMpumalanga
1.875
## provinceN.Cape
1.635
## provinceN.West
0.454
## provinceW.Cape
-0.796
## businesstypeRear livestock/poultry and sell e.g. chickens
```

```
-1.763
## businesstypeRender a professional service e.g. doctor, lawyer, accountant,
engineer, consultant             -1.484
## businesstypeRender a skilled service e.g. mechanic, plumber, hair salon,
barber, painting, landscaping      -2.651
## businesstypeRender building/construction services
1.959
## businesstypeRender other services e.g. car wash, garden services,
transport (taxi services), catering       -3.008
## businesstypeRender tourism-related services e.g.
accommodation/hotel/B&B/guest house, tour operators        -2.406
## businesstypeSell by-products of animals e.g. meat, eggs, milk
-0.920
## businesstypeSell something in the same form that I buy from someone else
(dont add value, e.g. cigarettes)  -1.863
## businesstypeSell something that I buy but add value to, e.g. repackage,
cook, etc                           -3.213
## businesstypeSell something that I collect from nature, e.g. herbs,
firewood, charcoal, thatch, sand, stone  -0.382
## businesstypeSell something that I get for free, e.g. second hand clothes,
scrap metal                        -1.144
## businesstypeSell something that I make e.g. crafts, clothes, furniture,
bricks                              -0.349
## wherebusoperateCar/truck/vehicle
0.665
## wherebusoperateDoor to door/Go to customers
-3.461
## wherebusoperateFarm/small holding
-2.246
## wherebusoperateOffice block/office park
-4.553
## wherebusoperateOnline - internet, phone selling
-3.710
## wherebusoperateOpen space-Isipingo
0.949
## wherebusoperateResidential premises - dwelling/garage/building on
residential premises                       -5.532
## wherebusoperateSchool Cafeteria
-2.131
## wherebusoperateShopping mall
-0.407
## wherebusoperateStall/table/container in a designated trading or market
area                                  -2.356
## wherebusoperateStreet/street corner/pavement
-5.199
## ownrentOwn the business premises
0.307
## tot_hours_wrk
0.231
## priv_indvDoesn't sell to Private Individuals
```

```
0.858
## other_small_busSells to other small enterpises
-1.050
## other_large_busSells to larger enterpises
1.280
## govSells to government
4.547
## tender_succNo
-0.711
## tender_succYes
1.907
## bank_loanYES
-1.523
## age
0.459
## highlevel_edu.L
4.381
## highlevel_edu.Q
-0.640
## highlevel_edu.C
-2.749
## highlevel_edu^4
-1.222
## highlevel_edu^5
6.884
## highlevel_edu^6
7.782
## highlevel_edu^7
7.320
## is_bus_only_src_incBusiness is the only source of income
0.615
## keep_fin_recKeep financial records
-0.026
## ovrall_fin_accessFormal
-1.555
## ovrall_fin_accessInformal
-2.281
## ovrall_fin_accessNot served
-2.303
## having_security_measuresHave security Measures
1.234
## suffered_crimeortheftBusiness suffered crime or theft in the last 12
months                                    -2.325
## is_bus_registeredBusiness is registered
0.386
## comp_fin_recsBusiness keeps computerized financial records
0.532
## owner_rentor_own_priv_resOwn the private residence
-3.568
## expto_custExport to outside of SA
```

```
-0.150
## supp_outof_saHave suppliers out of SA
0.215
## have_insHave Insurance
0.858
## off_good_credYes always
-0.644
## off_good_credYes, sometimes
0.935
## have_accesstohowmanybusfuncs
-1.723
## enterprise_classification.L
1.312
## enterprise_classification.Q
-0.134
## enterprise_classification.C
-1.940
## have_acces_to_howmany_ins_prod
-2.387
## source_ofcreditBorrowing from friends/family
-1.405
## source_ofcreditFormal
0.813
## source_ofcreditInformal
-0.531
## source_ofcreditNot served
-1.235
## age_of_business
1.962
## bus_turn_monthly_log
34.820
##
Pr(>|t|)
## (Intercept)
< 2e-16
## locationTribal area
0.330985
## locationUrban formal
0.009968
## locationUrban informal
0.033759
## provinceFree State
2.18e-05
## provinceGauteng
0.000587
## provinceKZN
1.25e-06
## provinceLimpopo
0.127921
## provinceMpumalanga
```

```
0.060891
## provinceN.Cape
0.102152
## provinceN.West
0.649963
## provinceW.Cape
0.425846
## businesstypeRear livestock/poultry and sell e.g. chickens
0.078006
## businesstypeRender a professional service e.g. doctor, lawyer, accountant,
engineer, consultant            0.137756
## businesstypeRender a skilled service e.g. mechanic, plumber, hair salon,
barber, painting, landscaping     0.008044
## businesstypeRender building/construction services
0.050176
## businesstypeRender other services e.g. car wash, garden services,
transport (taxi services), catering      0.002644
## businesstypeRender tourism-related services e.g.
accommodation/hotel/B&B/guest house, tour operators       0.016142
## businesstypeSell by-products of animals e.g. meat, eggs, milk
0.357613
## businesstypeSell something in the same form that I buy from someone else
(dont add value, e.g. cigarettes) 0.062538
## businesstypeSell something that I buy but add value to, e.g. repackage,
cook, etc                          0.001322
## businesstypeSell something that I collect from nature, e.g. herbs,
firewood, charcoal, thatch, sand, stone 0.702667
## businesstypeSell something that I get for free, e.g. second hand clothes,
scrap metal                       0.252833
## businesstypeSell something that I make e.g. crafts, clothes, furniture,
bricks                             0.727088
## wherebusoperateCar/truck/vehicle
0.505899
## wherebusoperateDoor to door/Go to customers
0.000542
## wherebusoperateFarm/small holding
0.024770
## wherebusoperateOffice block/office park
5.41e-06
## wherebusoperateOnline - internet, phone selling
0.000210
## wherebusoperateOpen space-Isipingo
0.342660
## wherebusoperateResidential premises - dwelling/garage/building on
residential premises                     3.31e-08
## wherebusoperateSchool Cafeteria
0.033127
## wherebusoperateShopping mall
0.684180
## wherebusoperateStall/table/container in a designated trading or market
```

```
area                                    0.018486
## wherebusoperateStreet/street corner/pavement
2.08e-07
## ownrentOwn the business premises
0.758497
## tot_hours_wrk
0.816999
## priv_indvDoesn't sell to Private Individuals
0.391025
## other_small_busSells to other small enterpises
0.293791
## other_large_busSells to larger enterpises
0.200692
## govSells to government
5.55e-06
## tender_succNo
0.477111
## tender_succYes
0.056582
## bank_loanYES
0.127705
## age
0.646027
## highlevel_edu.L
1.20e-05
## highlevel_edu.Q
0.521906
## highlevel_edu.C
0.005993
## highlevel_edu^4
0.221652
## highlevel_edu^5
6.47e-12
## highlevel_edu^6
8.41e-15
## highlevel_edu^7
2.82e-13
## is_bus_only_src_incBusiness is the only source of income
0.538309
## keep_fin_recKeep financial records
0.979290
## ovrall_fin_accessFormal
0.119919
## ovrall_fin_accessInformal
0.022615
## ovrall_fin_accessNot served
0.021296
## having_security_measuresHave security Measures
0.217134
## suffered_crimeortheftBusiness suffered crime or theft in the last 12
```

```
months                                    0.020132
## is_bus_registeredBusiness is registered
0.699180
## comp_fin_recsBusiness keeps computerized financial records
0.594631
## owner_rentor_own_priv_resOwn the private residence
0.000363
## expto_custExport to outside of SA
0.880403
## supp_outof_saHave suppliers out of SA
0.829982
## have_insHave Insurance
0.391100
## off_good_credYes always
0.519864
## off_good_credYes, sometimes
0.349752
## have_accesstohowmanybusfuncs
0.085026
## enterprise_classification.L
0.189409
## enterprise_classification.Q
0.893295
## enterprise_classification.C
0.052489
## have_acces_to_howmany_ins_prod
0.017012
## source_ofcreditBorrowing from friends/family
0.159962
## source_ofcreditFormal
0.416138
## source_ofcreditInformal
0.595513
## source_ofcreditNot served
0.217003
## age_of_business
0.049849
## bus_turn_monthly_log
< 2e-16
##
## (Intercept)
***
## locationTribal area
## locationUrban formal
**
## locationUrban informal
*
## provinceFree State
***
## provinceGauteng
```

```
***
## provinceKZN
***
## provinceLimpopo
## provinceMpumalanga
.
## provinceN.Cape
## provinceN.West
## provinceW.Cape
## businesstypeRear livestock/poultry and sell e.g. chickens
.
## businesstypeRender a professional service e.g. doctor, lawyer, accountant,
engineer, consultant
## businesstypeRender a skilled service e.g. mechanic, plumber, hair salon,
barber, painting, landscaping     **
## businesstypeRender building/construction services
.
## businesstypeRender other services e.g. car wash, garden services,
transport (taxi services), catering      **
## businesstypeRender tourism-related services e.g.
accommodation/hotel/B&B/guest house, tour operators       *
## businesstypeSell by-products of animals e.g. meat, eggs, milk
## businesstypeSell something in the same form that I buy from someone else
(dont add value, e.g. cigarettes) .
## businesstypeSell something that I buy but add value to, e.g. repackage,
cook, etc                          **
## businesstypeSell something that I collect from nature, e.g. herbs,
firewood, charcoal, thatch, sand, stone
## businesstypeSell something that I get for free, e.g. second hand clothes,
scrap metal
## businesstypeSell something that I make e.g. crafts, clothes, furniture,
bricks
## wherebusoperateCar/truck/vehicle
## wherebusoperateDoor to door/Go to customers
***
## wherebusoperateFarm/small holding
*
## wherebusoperateOffice block/office park
***
## wherebusoperateOnline - internet, phone selling
***
## wherebusoperateOpen space-Isipingo
## wherebusoperateResidential premises - dwelling/garage/building on
residential premises                     ***
## wherebusoperateSchool Cafeteria
*
## wherebusoperateShopping mall
## wherebusoperateStall/table/container in a designated trading or market
area                                *
## wherebusoperateStreet/street corner/pavement
```

```
***
## ownrentOwn the business premises
## tot_hours_wrk
## priv_indvDoesn't sell to Private Individuals
## other_small_busSells to other small enterpises
## other_large_busSells to larger enterpises
## govSells to government
***
## tender_succNo
## tender_succYes
.
## bank_loanYES
## age
## highlevel_edu.L
***
## highlevel_edu.Q
## highlevel_edu.C
**
## highlevel_edu^4
## highlevel_edu^5
***
## highlevel_edu^6
***
## highlevel_edu^7
***
## is_bus_only_src_incBusiness is the only source of income
## keep_fin_recKeep financial records
## ovrall_fin_accessFormal
## ovrall_fin_accessInformal
*
## ovrall_fin_accessNot served
*
## having_security_measuresHave security Measures
## suffered_crimeortheftBusiness suffered crime or theft in the last 12
months                                  *
## is_bus_registeredBusiness is registered
## comp_fin_recsBusiness keeps computerized financial records
## owner_rentor_own_priv_resOwn the private residence
***
## expto_custExport to outside of SA
## supp_outof_saHave suppliers out of SA
## have_insHave Insurance
## off_good_credYes always
## off_good_credYes, sometimes
## have_accesstohowmanybusfuncs
.
## enterprise_classification.L
## enterprise_classification.Q
## enterprise_classification.C
.
```

```
## have_acces_to_howmany_ins_prod
*
## source_ofcreditBorrowing from friends/family
## source_ofcreditFormal
## source_ofcreditInformal
## source_ofcreditNot served
## age_of_business
*
## bus_turn_monthly_log
***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.415 on 5598 degrees of freedom
## Multiple R-squared:  0.3154, Adjusted R-squared:  0.306
## F-statistic: 33.49 on 77 and 5598 DF,  p-value: < 2.2e-16
```

Dropping the revenue variable

From a couple of models and plots we have conducted thus far, its clear that business net revenue monthly is a robust predictor for monthly profits. However, revenue and profits typically measure the same thing, though the may be some differences in their values. Thus, revenue is not really a good predictor variable as it also logically relies on the same variables for it to be either big or small that monthly profits rely on, thus it doesn't really give us an inferential insight and it gives us a false sense of power into the factors that inform high business profits when it just measures the same thing, thus we will drop it from our model.

```
imputed_finscope_dataz <- subset(imputed_finscope_data4, select = -
c(bus_turn_monthly_log))
```

Model misspecification

Next we shift our attention to functional misspecification which is also another important issue. Functional misspecification is an issue where in our model specification we have failed to account for some important non-linear relationships leading to biases in the estimation of the population parameters or the statistics.

Below we run the Resettest to check for functional misspecification in the model. Note, this test does not tell us what form of misspecification we have, and it is generally hard to have a sense of the type of misspecification we are dealing with, thus it is rare the case that we are able to completely deal with it. The model runs a regression on y as the target variable and adds a slew of non-linear combinations of the fitted values on the x predictor space. The goal is to find if any form of misspecification exists. The reset test then uses the f-test to compare the original model with the modified model to see if the now-modified model performs better than the standard model.

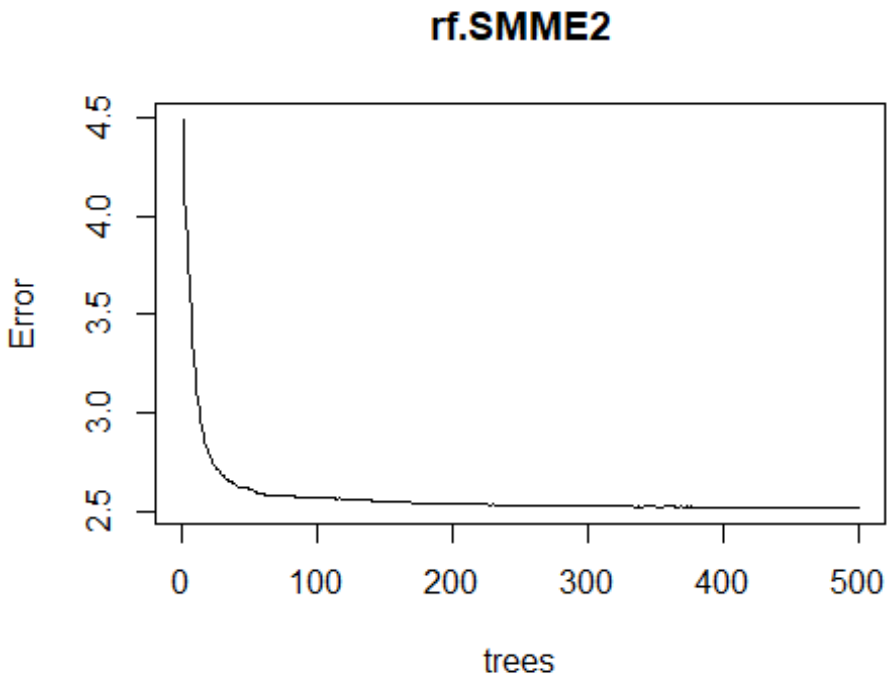
```
#Lets check for some functional misspecification..
resettest(model.log, power = 2:3, type = 'regressor',data =
imputed_finscope_data4)

##
##  RESET test
##
## data:  model.log
## RESET = 3.947, df1 = 12, df2 = 5586, p-value = 4.293e-06

#Ohk, it seems the p-value at 0.1013 is not sufficiently large enough.
#Meaning that the extended model with higher degree polynomials of the
explanatory variables is
#Not sufficiently different from the model without the higher degree
polynomial to warrant their inclusion..
#Thus, there is no sufficient evidence for functional misspecification..
```

From the test above, we see that indeed we do have some form of model misspecification. Meaning that the extended model with higher degree polynomials of the explanatory variables is sufficiently different from the model without the higher degree polynomial to warrant their inclusion. Thus, there is sufficient evidence for functional misspecification, meaning we might need to do some feature engineering on top of the ones we have made.

To achieve this, we can make use of the random forest algorithm to find one form of model misspecification, which is interaction, or variable that has some sort of predictive power when they interact with each other.

Random forest is a tree-based algorithm, that uses multiple trees instead of one to run regression and make predictions, and the prediction of those trees is averaged to produce one prediction. With random forests, the features that make up the tree are chosen at random at each iteration to boost performance.

```
set.seed(1234)

rf.SMME2 = randomForest(bus_net_profmnthly_log ~., data =
imputed_finscope_dataz, localImp = TRUE)

plot(rf.SMME2)
```



Interaction terms..

To find the interactions using the random forest algorithm, we make use of the *random forest explainer* package to find the interactions that have predictive power in the set of features we have.

The code below is used to get the minimal depth interactions, or rather interactions that are closest to the root node and thus have predictive power. Then the second batch of code shows the top interactions ordered by mean_min_depth which is the average distance of the interaction from the root node, where a small value of mean_min_depth is considered to be good compared to larger values.

```
importance_frame <- measure_importance(rf.SMME2)

(vars <- important_variables(importance_frame, k = 15, measures =
c("mean_min_depth", "no_of_trees")))
```

```
##  [1] "wherebusoperate"                "comp_fin_recs"
##  [3] "enterprise_classification"      "businesstype"
##  [5] "province"                       "highlevel_edu"
##  [7] "ovrall_fin_access"              "age"
##  [9] "age_of_business"                "having_security_measures"
## [11] "tot_hours_wrk"                  "is_bus_registered"
## [13] "source_ofcredit"                "have_accesstohowmanybusfuncs"
## [15] "have_acces_to_howmany_ins_prod"
```

```
interactions_frame2 <- min_depth_interactions(rf.SMME2, vars)
```

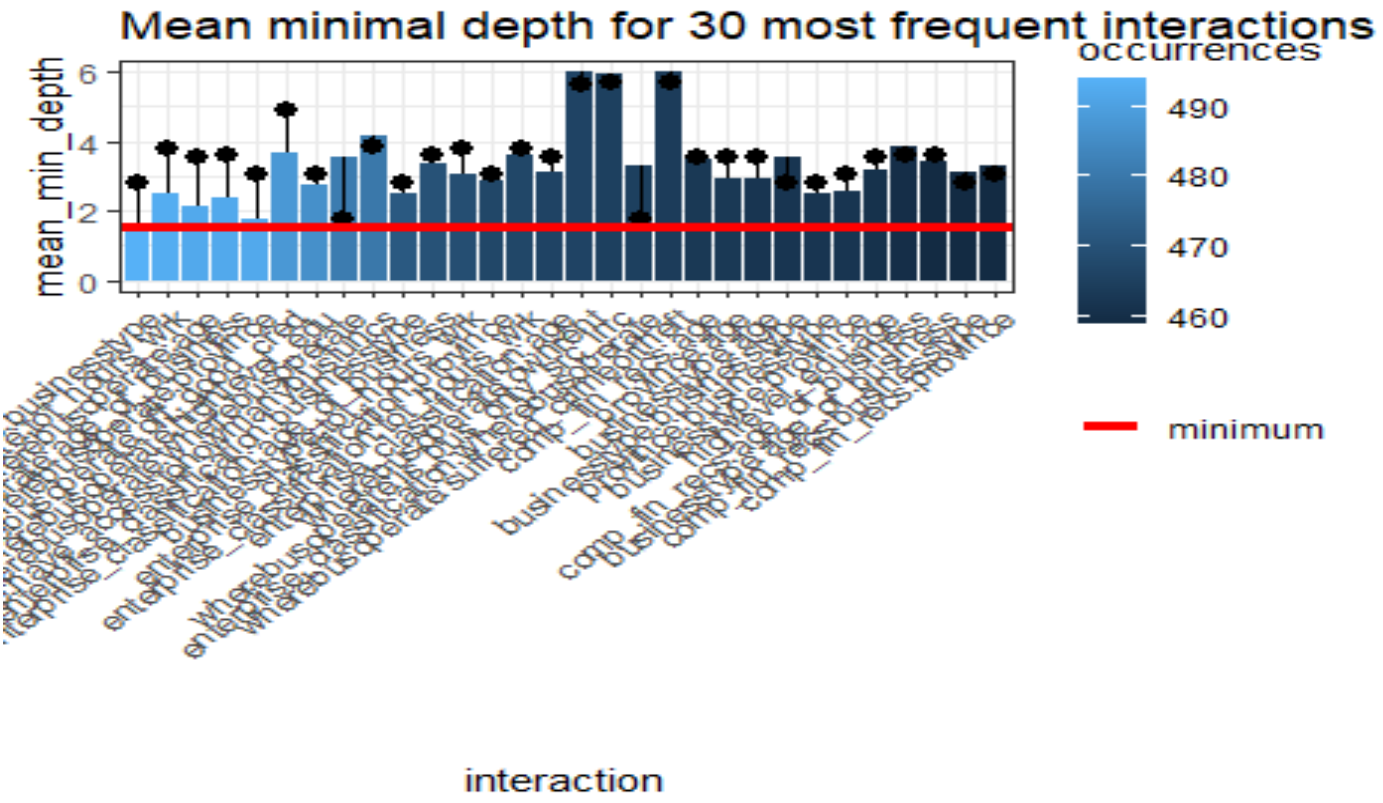
```
## Warning: `funs()` was deprecated in dplyr 0.8.0.
## ℹ Please use a list of either functions or lambdas:
##
## # Simple named list: list(mean = mean, median = median)
##
## # Auto named with `tibble::lst()`: tibble::lst(mean, median)
##
## # Using lambdas list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## ℹ The deprecated feature was likely used in the randomForestExplainer
package.
##   Please report the issue to the authors.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.

## Warning: There were 83327 warnings in `summarise()`.
## The first warning was:
## ℹ In argument: `wherebusoperate = min(wherebusoperate, na.rm = TRUE)`.
## ℹ In group 3: `tree = 1`, `split var = "bank_loan"`.
## Caused by warning in `min()`:
## ! no non-missing arguments to min; returning Inf
## ℹ Run `dplyr::last_dplyr_warnings()` to see the 83326 remaining warnings.
```

```
head(interactions_frame2[order(interactions_frame2$mean_min_depth, decreasing
= FALSE), ])
```

```
##            variable             root_variable mean_min_depth occurrences
## 60     businesstype           wherebusoperate       1.522267         494
## 375        province           wherebusoperate       1.781765         492
## 15              age           wherebusoperate       2.113749         492
## 30  age_of_business           wherebusoperate       2.387028         492
## 50     businesstype enterprise_classification       2.486275         472
## 57     businesstype                  province       2.504972         462
##                                interaction uncond_mean_min_depth
## 60           wherebusoperate:businesstype                 2.810
## 375              wherebusoperate:province                 3.040
## 15                    wherebusoperate:age                 3.572
## 30         wherebusoperate:age_of_business                 3.632
## 50  enterprise_classification:businesstype                 2.810
## 57                   province:businesstype                 2.810
```

Now, we plot the interactions frame of interaction and try to pick the most promising interactions to include into our final model.

```
plot_min_depth_interactions(interactions_frame2)
```



Nice, here we have a group of some important interactions on the basis of a random forest algorithm which used the greedy algorithm to decide on the split at each node, with features that can lead to the greatest reduction in error being more preferred. The importance of these nodes is based on how far are they from the initial node, meaning how much of the variation they account for. The closer each interaction is to the initial node, the more important such a node is and thus the more explanatory power it has. The plot uses a slew of metrics to decide on which interaction is important. One is the mean minimal depth, which measures the average distance of each interaction from the root node, with small mean minimal depths being preferred since this means a split at these nodes at the basis of the greedy algorithm would lead to the greatest improvement on the variability explained by the model. The second measure we look into pertains to the number of occurrences of each interaction in the random trees, with trees with the most occurrences being more important interactions. Occurrences are measured from the left to the right, with interactions in the left occurring the most and declining toward the left. The third measure looks into the unconditional mean depth, which looks into the average distance of the interaction from the root node across the trees in the ensemble, not just on the trees it forms part of. The unconditional mean depth is the most robust way to choose the interaction as some interactions may be important in only in the trees they appear in, but not overall, meaning they don't rank as high in the trees they are not part of.

It is in the balance of these three measures that we will choose the most important interactions, and also look into the cardinality of each of the interactions to ensure that we don't run into issues such as overfitting due to having categories that are too many and are

rare and thus do not generalize very well, or have too many categories which can lead to computational issues.

From the onset, we can see a couple of important interactions, with the most important being the interaction between where the business operates and the business type. This variable has the lowest mean minimal depth, a relatively small unconditional depth, and is second from the left meaning it has a relatively high occurrence rate. This tells us the profitability of certain business types can be contingent on their location, with certain locations being less favored. Then, on the same basis, we see where a business operates and the province is important, followed by where a business operates and its age, however, this has the potential to increase the cardinality of the data. The following interactions will also be considered, where a business operates and the highest level of education of the owner, business type and the province, location, and the province in which the business is located.

Creating the interaction variables

```
imputed_finscope_dataz$wherbusop_bustype <-
interaction(imputed_finscope_dataz$wherebusoperate,
imputed_finscope_dataz$businesstype)

imputed_finscope_dataz$wherbusop_prov <-
interaction(imputed_finscope_dataz$wherebusoperate,
imputed_finscope_dataz$province)


imputed_finscope_dataz$bustype_prov <-
interaction(imputed_finscope_dataz$province,
imputed_finscope_dataz$businesstype)
```

Model Fitting

Logged Regression model

Below, we fit a regression model using the *imputed_finscope data*. This will be the first model of the two we will fit, and we will use it as the baseline model. Then afterwards we fit the lasso regression model.

Note, that we previously transformed the y-variable by logging it to solve the issue of non-linearity and partially the issue of heteroscedasticity. To fit this model, we used the cross-validation method, where we separated the data into 10 folds, iteratively trained on the k-1 folds, and tested on the k-th fold at each iteration. Then, to find our coefficients and their associated statistical properties such as t-values, std errors, and p-values, we minimized the cross-validated error terms, to ensure that our model is able to generalize and not overfit. We then subsequently picked only the most significant variables at the 0.05 significant level, which is in line with the literature.

```
x = model.matrix(bus_net_profmnthly_log~., data = imputed_finscope_dataz)[, -
1] #The matrix of x-variables which we will use to fit the model.
```

```r
y = imputed_finscope_dataz$bus_net_profmnthly_log

trControl <- trainControl(method = 'cv', number = 10)

model <- train(x, y, method = 'lm', trControl = trControl)

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient
fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient
fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient
fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient
fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient
fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient
fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient
fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient
fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient
fit
## may be misleading

## Warning in predict.lm(modelFit, newdata): prediction from a rank-deficient
fit
## may be misleading

summary_lm_cv  <- summary(model)
```

```r
#Get the coefficients we need and store them in the coef_lm1 variable
coef_lm1 <- summary_lm_cv$coefficients


#Pick the most important variables at a 0.05 significance level
sig_coef_lm_cv <- coef_lm1[coef_lm1[, 4] <= 0.05, ]


print(sig_coef_lm_cv)
```

```
##
Estimate
## (Intercept)
8.652695732
## `locationUrban formal`
0.274845598
## `locationUrban informal`
0.263561750
## `wherebusoperateOpen space-Isipingo`
3.532479770
## tot_hours_wrk
0.015888976
## `priv_indvDoesn't sell to Private Individuals`
0.565945268
## `other_large_busSells to larger enterpises`
0.263651421
## `govSells to government`
0.638681996
## tender_succNo
-0.360299151
## highlevel_edu.L
0.725839829
## highlevel_edu.C
-0.254451860
## `highlevel_edu^5`
0.488954962
## `highlevel_edu^6`
0.729536947
## `highlevel_edu^7`
0.365322321
## `ovrall_fin_accessNot served`
-0.166424992
## `comp_fin_recsBusiness keeps computerized financial records`
0.235240449
## `owner_rentor_own_priv_resOwn the private residence`
-0.203355751
## enterprise_classification.L
0.971795298
## `source_ofcreditBorrowing from friends/family`
-0.396289550
```

```
## age_of_business
0.008561863
## `wherbusop_bustypeStall/table/container in a designated trading or market
area.Render other services e.g. car wash, garden services, transport (taxi
services), catering` -2.144740939
## `wherbusop_bustypeBusiness park/Premises dedicated to my business -
hotel/accommodation facility/factory/workshop.Sell by-products of animals
e.g. meat, eggs, milk`       2.288683461
## `wherbusop_bustypeStall/table/container in a designated trading or market
area.Sell by-products of animals e.g. meat, eggs, milk`
2.674490252
## `wherbusop_bustypeFarm/small holding.Sell something that I buy but add
value to, e.g. repackage, cook, etc`
-2.377861932
## `wherbusop_provFarm/small holding.E.Cape`
2.636624413
## `wherbusop_provFarm/small holding.Gauteng`
1.888319827
## `wherbusop_provFarm/small holding.KZN`
1.827805513
## `wherbusop_provFarm/small holding.Limpopo`
2.690566812
## `wherbusop_provFarm/small holding.N.Cape`
1.837757795
## `bustype_provMpumalanga.Grow something and sell, e.g. fruit, vegetables,
plants (like a nursery)`
-1.409463124
## `bustype_provKZN.Render other services e.g. car wash, garden services,
transport (taxi services), catering`
-0.883336454
## `bustype_provMpumalanga.Render other services e.g. car wash, garden
services, transport (taxi services), catering`
-1.226478568
## `bustype_provKZN.Sell something in the same form that I buy from someone
else (dont add value, e.g. cigarettes)`
-0.734470254
##
Std. Error
## (Intercept)
1.294060538
## `locationUrban formal`
0.086447849
## `locationUrban informal`
0.109677804
## `wherebusoperateOpen space-Isipingo`
1.801656320
## tot_hours_wrk
0.007044594
## `priv_indvDoesn't sell to Private Individuals`
0.170407801
```

```
## `other_large_busSells to larger enterpises`
0.134239176
## `govSells to government`
0.157536698
## tender_succNo
0.153574205
## highlevel_edu.L
0.137314952
## highlevel_edu.C
0.116138889
## `highlevel_edu^5`
0.092004406
## `highlevel_edu^6`
0.107929038
## `highlevel_edu^7`
0.073256800
## `ovrall_fin_accessNot served`
0.056052953
## `comp_fin_recsBusiness keeps computerized financial records`
0.105268015
## `owner_rentor_own_priv_resOwn the private residence`
0.069828387
## enterprise_classification.L
0.397628527
## `source_ofcreditBorrowing from friends/family`
0.168446065
## age_of_business
0.003293980
## `wherbusop_bustypeStall/table/container in a designated trading or market
area.Render other services e.g. car wash, garden services, transport (taxi
services), catering` 1.076178248
## `wherbusop_bustypeBusiness park/Premises dedicated to my business -
hotel/accommodation facility/factory/workshop.Sell by-products of animals
e.g. meat, eggs, milk`      1.140444185
## `wherbusop_bustypeStall/table/container in a designated trading or market
area.Sell by-products of animals e.g. meat, eggs, milk`
1.085021974
## `wherbusop_bustypeFarm/small holding.Sell something that I buy but add
value to, e.g. repackage, cook, etc`
0.903807214
## `wherbusop_provFarm/small holding.E.Cape`
0.866905740
## `wherbusop_provFarm/small holding.Gauteng`
0.708581007
## `wherbusop_provFarm/small holding.KZN`
0.886652736
## `wherbusop_provFarm/small holding.Limpopo`
0.700948487
## `wherbusop_provFarm/small holding.N.Cape`
0.731193269
```

```
## `bustype_provMpumalanga.Grow something and sell, e.g. fruit, vegetables,
plants (like a nursery)`
0.652451601
## `bustype_provKZN.Render other services e.g. car wash, garden services,
transport (taxi services), catering`
0.446764173
## `bustype_provMpumalanga.Render other services e.g. car wash, garden
services, transport (taxi services), catering`
0.623887366
## `bustype_provKZN.Sell something in the same form that I buy from someone
else (dont add value, e.g. cigarettes)`
0.322393539
##
t value
## (Intercept)
6.686469
## `locationUrban formal`
3.179323
## `locationUrban informal`
2.403055
## `wherebusoperateOpen space-Isipingo`
1.960685
## tot_hours_wrk
2.255485
## `priv_indvDoesn't sell to Private Individuals`
3.321123
## `other_large_busSells to larger enterpises`
1.964042
## `govSells to government`
4.054179
## tender_succNo
-2.346092
## highlevel_edu.L
5.285949
## highlevel_edu.C
-2.190927
## `highlevel_edu^5`
5.314473
## `highlevel_edu^6`
6.759413
## `highlevel_edu^7`
4.986872
## `ovrall_fin_accessNot served`
-2.969067
## `comp_fin_recsBusiness keeps computerized financial records`
2.234681
## `owner_rentor_own_priv_resOwn the private residence`
-2.912222
## enterprise_classification.L
2.443978
```

```
## `source_ofcreditBorrowing from friends/family`
-2.352620
## age_of_business
2.599245
## `wherbusop_bustypeStall/table/container in a designated trading or market
area.Render other services e.g. car wash, garden services, transport (taxi
services), catering` -1.992924
## `wherbusop_bustypeBusiness park/Premises dedicated to my business -
hotel/accommodation facility/factory/workshop.Sell by-products of animals
e.g. meat, eggs, milk`       2.006835
## `wherbusop_bustypeStall/table/container in a designated trading or market
area.Sell by-products of animals e.g. meat, eggs, milk`
2.464918
## `wherbusop_bustypeFarm/small holding.Sell something that I buy but add
value to, e.g. repackage, cook, etc`
-2.630939
## `wherbusop_provFarm/small holding.E.Cape`
3.041420
## `wherbusop_provFarm/small holding.Gauteng`
2.664931
## `wherbusop_provFarm/small holding.KZN`
2.061467
## `wherbusop_provFarm/small holding.Limpopo`
3.838466
## `wherbusop_provFarm/small holding.N.Cape`
2.513368
## `bustype_provMpumalanga.Grow something and sell, e.g. fruit, vegetables,
plants (like a nursery)`
-2.160257
## `bustype_provKZN.Render other services e.g. car wash, garden services,
transport (taxi services), catering`
-1.977187
## `bustype_provMpumalanga.Render other services e.g. car wash, garden
services, transport (taxi services), catering`
-1.965865
## `bustype_provKZN.Sell something in the same form that I buy from someone
else (dont add value, e.g. cigarettes)`
-2.278179
##
Pr(>|t|)
## (Intercept)
2.519138e-11
## `locationUrban formal`
1.484618e-03
## `locationUrban informal`
1.629260e-02
## `wherebusoperateOpen space-Isipingo`
4.996765e-02
## tot_hours_wrk
2.414316e-02
```

```
## `priv_indvDoesn't sell to Private Individuals`
9.025709e-04
## `other_large_busSells to larger enterpises`
4.957689e-02
## `govSells to government`
5.102523e-05
## tender_succNo
1.900776e-02
## highlevel_edu.L
1.300193e-07
## highlevel_edu.C
2.850005e-02
## `highlevel_edu^5`
1.113037e-07
## `highlevel_edu^6`
1.533117e-11
## `highlevel_edu^7`
6.330297e-07
## `ovrall_fin_accessNot served`
3.000307e-03
## `comp_fin_recsBusiness keeps computerized financial records`
2.547943e-02
## `owner_rentor_own_priv_resOwn the private residence`
3.603520e-03
## enterprise_classification.L
1.455840e-02
## `source_ofcreditBorrowing from friends/family`
1.867771e-02
## age_of_business
9.368555e-03
## `wherbusop_bustypeStall/table/container in a designated trading or market
area.Render other services e.g. car wash, garden services, transport (taxi
services), catering` 4.632054e-02
## `wherbusop_bustypeBusiness park/Premises dedicated to my business -
hotel/accommodation facility/factory/workshop.Sell by-products of animals
e.g. meat, eggs, milk`      4.481745e-02
## `wherbusop_bustypeStall/table/container in a designated trading or market
area.Sell by-products of animals e.g. meat, eggs, milk`
1.373565e-02
## `wherbusop_bustypeFarm/small holding.Sell something that I buy but add
value to, e.g. repackage, cook, etc`
8.539344e-03
## `wherbusop_provFarm/small holding.E.Cape`
2.366063e-03
## `wherbusop_provFarm/small holding.Gauteng`
7.723523e-03
## `wherbusop_provFarm/small holding.KZN`
3.930670e-02
## `wherbusop_provFarm/small holding.Limpopo`
1.252340e-04
```

```
## `wherbusop_provFarm/small holding.N.Cape`
1.198760e-02
## `bustype_provMpumalanga.Grow something and sell, e.g. fruit, vegetables,
plants (like a nursery)`
3.079706e-02
## `bustype_provKZN.Render other services e.g. car wash, garden services,
transport (taxi services), catering`
4.807173e-02
## `bustype_provMpumalanga.Render other services e.g. car wash, garden
services, transport (taxi services), catering`
4.936581e-02
## `bustype_provKZN.Sell something in the same form that I buy from someone
else (dont add value, e.g. cigarettes)`
2.275517e-02
```

The Lasso regression

In this final phase of our analysis, we apply the lasso regression, which is the normal regression model but with a penalization term. The lasso regression acts as both a feature selection method and a regularization method which will aid us in enhancing the inferential power of our model. We will use this model for the purposes of feature selection and thus be able to get the features that are most important in explaining profitability. With lasso regression, we add a penalty term to the usual ordinarily least squares formulation which seeks to find a model that minimizes the distance between the target variable and the model-predicted values. What the lasso does is add a penalty for overfitting with a regularization term lambda. What this term ensures is that the model does not overfit the data by adding a bias into the data, to avoid overconfidence, especially in the instance where we have a small dataset.

Because we add a penalty term, to still be able to minimize the squared difference between the predicted and actual term, the model will be forced to shrink the coefficients of the various features in the model. However, some of the features are not that important and have a small weight in modeling y, thus the lasso model will allow for these to go all the way to zero, and thus fall away from the model, unlike other regularization models which do not allow for dropping of variables. This is done to balance the the stress that the regularization term has on the model which increases the sum of squared errors, and thus to mitigate that and still this sse value low, the model decreases the coefficients of the explainer variables.

However, since in this analysis, we are interested in the most robust relationships with y and thus we will use the lasso to drop some of the not-so-important variables. Here, in the code below, we extract the matrix of data that contains the feature variables and assign them to X, and we also extract the variable y which is our target and we assign it to the value y.

```
library(glmnet) #The package we will use to fit the L1/L2 functions...
```

```
## Warning: package 'glmnet' was built under R version 4.2.3

## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack

## Loaded glmnet 4.1-7

#This package does not use formulae
x = model.matrix(bus_net_profmnthly_log~., data = imputed_finscope_dataz)[, -
1] #The matrix of x-variables which we will use to fit the model.
y = imputed_finscope_dataz$bus_net_profmnthly_log #The target variable
```

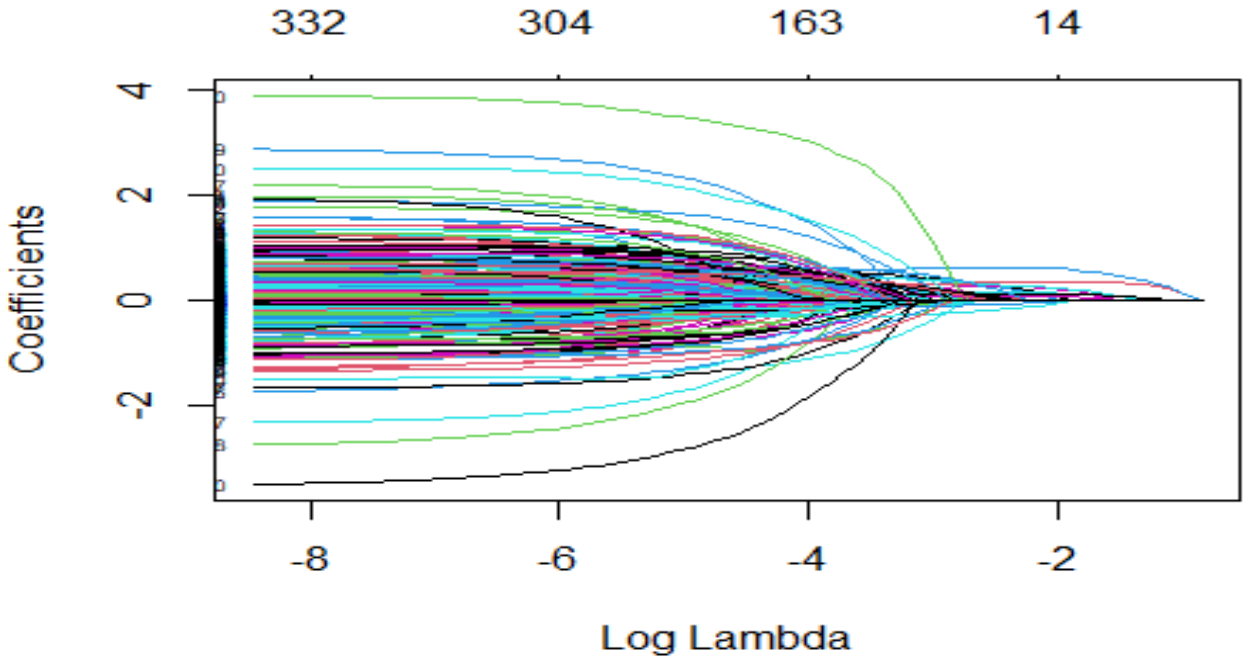
Here we fit the model using the imputed data. We use the extracted y and x from above.

```
fit.lasso = glmnet(x,y, data = imputed_finscope_dataz ) #The default is 1,
which is the lasso
```

Below we plot the model we have fitted above. On the x-axis, we have the log of the regularization term lambda. It decreases from the extreme negatives towards 0 and the positive side. On the y-axis, we have the values of the coefficients, and we have in the plot for each color the different features we have in our model, and their corresponding values as the log of the regularization term changes. On top of the model, we have a number of features we have in the model.

The key insight from this model is the fact that as the log of lambda increases, the coefficients of the various features in the model also tend to decline and at some point with a very large value of lambda, they become zero. Also, the log of lambda is very small, the penalty term is negligible, and thus the model is analogous to an ordinary least squares model. Also, the number of variables in the model also tends to decline as the value of the log of lambda increases. To find the optimal value of lambda that balances the performance of our model and the number of variables, we will use cross-validation.

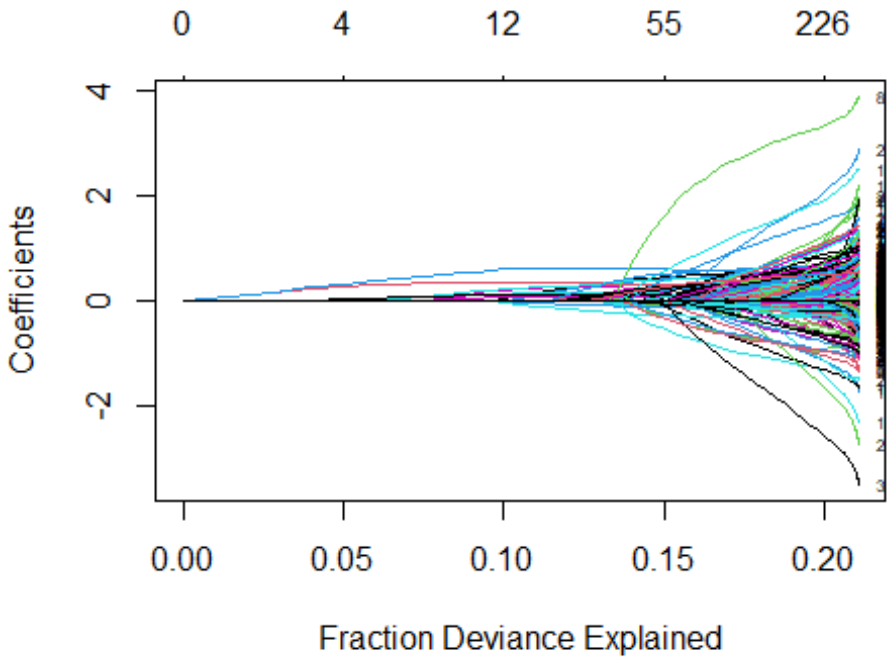
```
plot(fit.lasso, xvar = 'lambda', label= TRUE)
```



The plot below also enhances the model plot and explains how the deviation of our model prediction from the actual value relative to the mean prediction deviation from the actual value is improved by the increase in the number of variables in the model. Thus, on the x-axis, we have the deviation explained by our model relative to the mean model, and on the y-axis, we have the coefficient values. Initially, we have a few lines sprung up, meaning the penalty term is very high and thus there are a few features included in the model that are not suppressed by the penalty term. This entails that there's little variation being explained by our model. As the penalty term is relaxed and lambda becomes smaller, the variables included in the model increase as seen by the colored lines springing up.

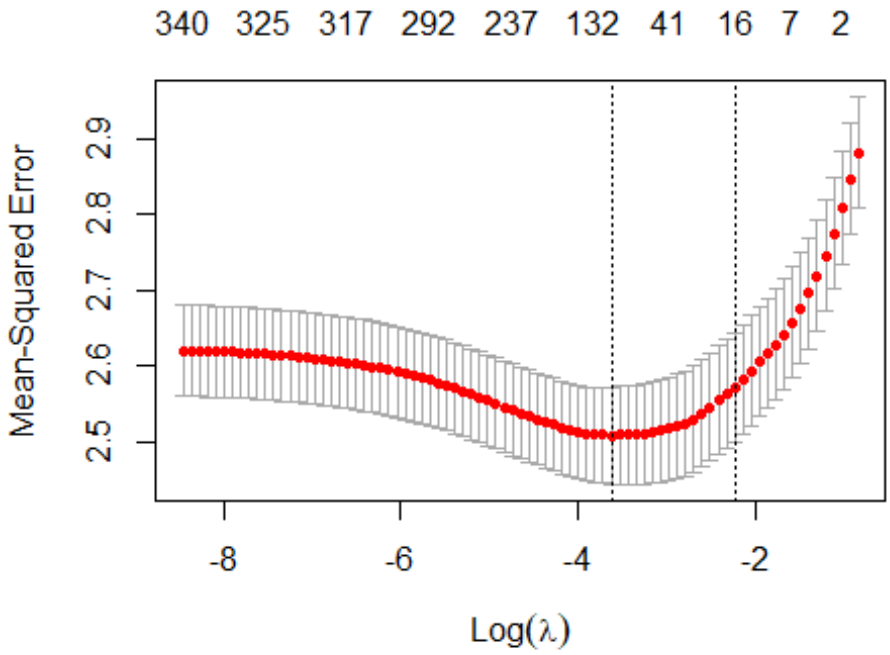
When the penalty term is very high and the number of variables included in the model is low, the variance explained is low, and it gradually improves as variables are added and the penalty term is relaxed. There's a general improvement in the deviance explained by our model up until say 0.05 where there are 12 variables. At around 0.15, there's an explosion of variables included in the model, all the way to around 0.20, which might be signal overfitting of our model by including too many features for such a small improvement in variability/deviance explained. Again, to find the balance between overfitting and the appropriate number of variables included, we will need to use cross-validation.

```
plot(fit.lasso, xvar = 'dev', label= TRUE)
```



Below we now run the cross-validation procedure to find the optimal regularization parameter value. The CV procedure works this way, instead of having a training and a validation dataset, we iteratively demarcate a portion of the data as the test set and use the rest as the training set. For instance, you divide your data into k-folds, then use k-1 folds to train, and then test on the k-th fold. Then you repeat the process but pick a different fold to test on each time. Then, use the procedure is to find the model that minimizes the cross-validated error term and simultaneously find the lambda that minimizes the cross-validated error term.

```
cv.lasso = cv.glmnet(x,y)

plot(cv.lasso)
```



From the graph above, it is clear where the log of lambda is at, this is the point where the mean squared error of the cross-validated error term is minimized.

Below, we run the final model, which has been cross-validated

```
coef(cv.lasso)

## 455 x 1 sparse Matrix of class "dgCMatrix"
##
s1
## (Intercept)
7.804484535
## locationTribal area
.
## locationUrban formal
0.031609977
## locationUrban informal
.
## provinceFree State
.
## provinceGauteng
.
## provinceKZN
.
## provinceLimpopo
.
## provinceMpumalanga
.
```

```
## provinceN.Cape
.
## provinceN.West
.
## provinceW.Cape
.
## businesstypeRear livestock/poultry and sell e.g. chickens
.
## businesstypeRender a professional service e.g. doctor, lawyer, accountant,
engineer, consultant
.
## businesstypeRender a skilled service e.g. mechanic, plumber, hair salon,
barber, painting, landscaping
.
## businesstypeRender building/construction services
.
## businesstypeRender other services e.g. car wash, garden services,
transport (taxi services), catering
.
## businesstypeRender tourism-related services e.g.
accommodation/hotel/B&B/guest house, tour operators
.
## businesstypeSell by-products of animals e.g. meat, eggs, milk
.
## businesstypeSell something in the same form that I buy from someone else
(dont add value, e.g. cigarettes)
.
## businesstypeSell something that I buy but add value to, e.g. repackage,
cook, etc
.
## businesstypeSell something that I collect from nature, e.g. herbs,
firewood, charcoal, thatch, sand, stone
.
## businesstypeSell something that I get for free, e.g. second hand clothes,
scrap metal
.
## businesstypeSell something that I make e.g. crafts, clothes, furniture,
bricks
.
## wherebusoperateCar/truck/vehicle
.
## wherebusoperateDoor to door/Go to customers
.
## wherebusoperateFarm/small holding
.
## wherebusoperateOffice block/office park
.
## wherebusoperateOnline - internet, phone selling
.
## wherebusoperateOpen space-Isipingo
```

```
.
## wherebusoperateResidential premises - dwelling/garage/building on
residential premises
-0.035514977
## wherebusoperateSchool Cafeteria
.
## wherebusoperateShopping mall
.
## wherebusoperateStall/table/container in a designated trading or market
area
.
## wherebusoperateStreet/street corner/pavement
.
## ownrentOwn the business premises
.
## tot_hours_wrk
.
## priv_indvDoesn't sell to Private Individuals
0.164180000
## other_small_busSells to other small enterpises
.
## other_large_busSells to larger enterpises
0.268356526
## govSells to government
0.216646880
## tender_succNo
.
## tender_succYes
.
## bank_loanYES
.
## age
.
## highlevel_edu.L
0.280646618
## highlevel_edu.Q
.
## highlevel_edu.C
.
## highlevel_edu^4
.
## highlevel_edu^5
.
## highlevel_edu^6
.
## highlevel_edu^7
.
## is_bus_only_src_incBusiness is the only source of income
.
## keep_fin_recKeep financial records
```

```
.
## ovrall_fin_accessFormal
.
## ovrall_fin_accessInformal
.
## ovrall_fin_accessNot served
-0.038535276
## having_security_measuresHave security Measures
0.162793109
## suffered_crimeortheftBusiness suffered crime or theft in the last 12
months
.
## is_bus_registeredBusiness is registered
0.147557193
## comp_fin_recsBusiness keeps computerized financial records
0.369624353
## owner_rentor_own_priv_resOwn the private residence
.
## expto_custExport to outside of SA
.
## supp_outof_saHave suppliers out of SA
.
## have_insHave Insurance
0.103045068
## off_good_credYes always
.
## off_good_credYes, sometimes
.
## have_accesstohowmanybusfuncs
0.007120513
## enterprise_classification.L
0.641279600
## enterprise_classification.Q
.
## enterprise_classification.C
.
## have_acces_to_howmany_ins_prod
0.002553780
## source_ofcreditBorrowing from friends/family
.
## source_ofcreditFormal
.
## source_ofcreditInformal
.
## source_ofcreditNot served
.
## age_of_business
.
## wherbusop_bustypeCar/truck/vehicle.Grow something and sell, e.g. fruit,
vegetables, plants (like a nursery)
```

.
wherbusop_bustypeDoor to door/Go to customers.Grow something and sell, e.g. fruit, vegetables, plants (like a nursery)
.
wherbusop_bustypeFarm/small holding.Grow something and sell, e.g. fruit, vegetables, plants (like a nursery)
.
wherbusop_bustypeOffice block/office park.Grow something and sell, e.g. fruit, vegetables, plants (like a nursery)
.
wherbusop_bustypeOnline - internet, phone selling.Grow something and sell, e.g. fruit, vegetables, plants (like a nursery)
.
wherbusop_bustypeOpen space-Isipingo.Grow something and sell, e.g. fruit, vegetables, plants (like a nursery)
.
wherbusop_bustypeResidential premises - dwelling/garage/building on residential premises.Grow something and sell, e.g. fruit, vegetables, plants (like a nursery) .
wherbusop_bustypeSchool Cafeteria.Grow something and sell, e.g. fruit, vegetables, plants (like a nursery)
.
wherbusop_bustypeShopping mall.Grow something and sell, e.g. fruit, vegetables, plants (like a nursery)
.
wherbusop_bustypeStall/table/container in a designated trading or market area.Grow something and sell, e.g. fruit, vegetables, plants (like a nursery)
.
wherbusop_bustypeStreet/street corner/pavement.Grow something and sell, e.g. fruit, vegetables, plants (like a nursery)
.
wherbusop_bustypeBusiness park/Premises dedicated to my business - hotel/accommodation facility/factory/workshop.Rear livestock/poultry and sell e.g. chickens .
wherbusop_bustypeCar/truck/vehicle.Rear livestock/poultry and sell e.g. chickens
.
wherbusop_bustypeDoor to door/Go to customers.Rear livestock/poultry and sell e.g. chickens
.
wherbusop_bustypeFarm/small holding.Rear livestock/poultry and sell e.g. chickens
.
wherbusop_bustypeOffice block/office park.Rear livestock/poultry and sell e.g. chickens
.
wherbusop_bustypeOnline - internet, phone selling.Rear livestock/poultry and sell e.g. chickens
.
wherbusop_bustypeOpen space-Isipingo.Rear livestock/poultry and sell e.g.

chickens
.
wherbusop_bustypeResidential premises - dwelling/garage/building on residential premises.Rear livestock/poultry and sell e.g. chickens
.
wherbusop_bustypeSchool Cafeteria.Rear livestock/poultry and sell e.g. chickens
.
wherbusop_bustypeShopping mall.Rear livestock/poultry and sell e.g. chickens
.
wherbusop_bustypeStall/table/container in a designated trading or market area.Rear livestock/poultry and sell e.g. chickens
.
wherbusop_bustypeStreet/street corner/pavement.Rear livestock/poultry and sell e.g. chickens
.
wherbusop_bustypeBusiness park/Premises dedicated to my business - hotel/accommodation facility/factory/workshop.Render a professional service e.g. doctor, lawyer, accountant, engineer, consultant .
wherbusop_bustypeCar/truck/vehicle.Render a professional service e.g. doctor, lawyer, accountant, engineer, consultant
.
wherbusop_bustypeDoor to door/Go to customers.Render a professional service e.g. doctor, lawyer, accountant, engineer, consultant
.
wherbusop_bustypeFarm/small holding.Render a professional service e.g. doctor, lawyer, accountant, engineer, consultant
.
wherbusop_bustypeOffice block/office park.Render a professional service e.g. doctor, lawyer, accountant, engineer, consultant
.
wherbusop_bustypeOnline - internet, phone selling.Render a professional service e.g. doctor, lawyer, accountant, engineer, consultant
.
wherbusop_bustypeOpen space-Isipingo.Render a professional service e.g. doctor, lawyer, accountant, engineer, consultant
.
wherbusop_bustypeResidential premises - dwelling/garage/building on residential premises.Render a professional service e.g. doctor, lawyer, accountant, engineer, consultant .
wherbusop_bustypeSchool Cafeteria.Render a professional service e.g. doctor, lawyer, accountant, engineer, consultant
.
wherbusop_bustypeShopping mall.Render a professional service e.g. doctor, lawyer, accountant, engineer, consultant
.
wherbusop_bustypeStall/table/container in a designated trading or market area.Render a professional service e.g. doctor, lawyer, accountant, engineer, consultant .

```
## wherbusop_bustypeStreet/street corner/pavement.Render a professional
service e.g. doctor, lawyer, accountant, engineer, consultant
.
## wherbusop_bustypeBusiness park/Premises dedicated to my business -
hotel/accommodation facility/factory/workshop.Render a skilled service e.g.
mechanic, plumber, hair salon, barber, painting, landscaping      .
## wherbusop_bustypeCar/truck/vehicle.Render a skilled service e.g. mechanic,
plumber, hair salon, barber, painting, landscaping
.
## wherbusop_bustypeDoor to door/Go to customers.Render a skilled service
e.g. mechanic, plumber, hair salon, barber, painting, landscaping
.
## wherbusop_bustypeFarm/small holding.Render a skilled service e.g.
mechanic, plumber, hair salon, barber, painting, landscaping
.
## wherbusop_bustypeOffice block/office park.Render a skilled service e.g.
mechanic, plumber, hair salon, barber, painting, landscaping
.
## wherbusop_bustypeOnline - internet, phone selling.Render a skilled service
e.g. mechanic, plumber, hair salon, barber, painting, landscaping
.
## wherbusop_bustypeOpen space-Isipingo.Render a skilled service e.g.
mechanic, plumber, hair salon, barber, painting, landscaping
.
## wherbusop_bustypeResidential premises - dwelling/garage/building on
residential premises.Render a skilled service e.g. mechanic, plumber, hair
salon, barber, painting, landscaping                                  .
## wherbusop_bustypeSchool Cafeteria.Render a skilled service e.g. mechanic,
plumber, hair salon, barber, painting, landscaping
.
## wherbusop_bustypeShopping mall.Render a skilled service e.g. mechanic,
plumber, hair salon, barber, painting, landscaping
.
## wherbusop_bustypeStall/table/container in a designated trading or market
area.Render a skilled service e.g. mechanic, plumber, hair salon, barber,
painting, landscaping                                              .
## wherbusop_bustypeStreet/street corner/pavement.Render a skilled service
e.g. mechanic, plumber, hair salon, barber, painting, landscaping
.
## wherbusop_bustypeBusiness park/Premises dedicated to my business -
hotel/accommodation facility/factory/workshop.Render building/construction
services                                                            .
## wherbusop_bustypeCar/truck/vehicle.Render building/construction services
.
## wherbusop_bustypeDoor to door/Go to customers.Render building/construction
services
.
## wherbusop_bustypeFarm/small holding.Render building/construction services
.
## wherbusop_bustypeOffice block/office park.Render building/construction
```

```
services
.
## wherbusop_bustypeOnline - internet, phone selling.Render
building/construction services
.
## wherbusop_bustypeOpen space-Isipingo.Render building/construction services
.
## wherbusop_bustypeResidential premises - dwelling/garage/building on
residential premises.Render building/construction services
.
## wherbusop_bustypeSchool Cafeteria.Render building/construction services
.
## wherbusop_bustypeShopping mall.Render building/construction services
.
## wherbusop_bustypeStall/table/container in a designated trading or market
area.Render building/construction services
.
## wherbusop_bustypeStreet/street corner/pavement.Render
building/construction services
.
## wherbusop_bustypeBusiness park/Premises dedicated to my business -
hotel/accommodation facility/factory/workshop.Render other services e.g. car
wash, garden services, transport (taxi services), catering       .
## wherbusop_bustypeCar/truck/vehicle.Render other services e.g. car wash,
garden services, transport (taxi services), catering
.
## wherbusop_bustypeDoor to door/Go to customers.Render other services e.g.
car wash, garden services, transport (taxi services), catering
.
## wherbusop_bustypeFarm/small holding.Render other services e.g. car wash,
garden services, transport (taxi services), catering
.
## wherbusop_bustypeOffice block/office park.Render other services e.g. car
wash, garden services, transport (taxi services), catering
.
## wherbusop_bustypeOnline - internet, phone selling.Render other services
e.g. car wash, garden services, transport (taxi services), catering
.
## wherbusop_bustypeOpen space-Isipingo.Render other services e.g. car wash,
garden services, transport (taxi services), catering
.
## wherbusop_bustypeResidential premises - dwelling/garage/building on
residential premises.Render other services e.g. car wash, garden services,
transport (taxi services), catering                                  .
## wherbusop_bustypeSchool Cafeteria.Render other services e.g. car wash,
garden services, transport (taxi services), catering
.
## wherbusop_bustypeShopping mall.Render other services e.g. car wash, garden
services, transport (taxi services), catering
.
```

```
## wherbusop_bustypeStall/table/container in a designated trading or market
area.Render other services e.g. car wash, garden services, transport (taxi
services), catering                                                    .
## wherbusop_bustypeStreet/street corner/pavement.Render other services e.g.
car wash, garden services, transport (taxi services), catering
.
## wherbusop_bustypeBusiness park/Premises dedicated to my business -
hotel/accommodation facility/factory/workshop.Render tourism-related services
e.g. accommodation/hotel/B&B/guest house, tour operators        .
## wherbusop_bustypeCar/truck/vehicle.Render tourism-related services e.g.
accommodation/hotel/B&B/guest house, tour operators
.
## wherbusop_bustypeDoor to door/Go to customers.Render tourism-related
services e.g. accommodation/hotel/B&B/guest house, tour operators
.
## wherbusop_bustypeFarm/small holding.Render tourism-related services e.g.
accommodation/hotel/B&B/guest house, tour operators
.
## wherbusop_bustypeOffice block/office park.Render tourism-related services
e.g. accommodation/hotel/B&B/guest house, tour operators
.
## wherbusop_bustypeOnline - internet, phone selling.Render tourism-related
services e.g. accommodation/hotel/B&B/guest house, tour operators
.
## wherbusop_bustypeOpen space-Isipingo.Render tourism-related services e.g.
accommodation/hotel/B&B/guest house, tour operators
.
## wherbusop_bustypeResidential premises - dwelling/garage/building on
residential premises.Render tourism-related services e.g.
accommodation/hotel/B&B/guest house, tour operators
.
## wherbusop_bustypeSchool Cafeteria.Render tourism-related services e.g.
accommodation/hotel/B&B/guest house, tour operators
.
## wherbusop_bustypeShopping mall.Render tourism-related services e.g.
accommodation/hotel/B&B/guest house, tour operators
.
## wherbusop_bustypeStall/table/container in a designated trading or market
area.Render tourism-related services e.g. accommodation/hotel/B&B/guest
house, tour operators                                                  .
## wherbusop_bustypeStreet/street corner/pavement.Render tourism-related
services e.g. accommodation/hotel/B&B/guest house, tour operators
.
## wherbusop_bustypeBusiness park/Premises dedicated to my business -
hotel/accommodation facility/factory/workshop.Sell by-products of animals
e.g. meat, eggs, milk                                                  .
## wherbusop_bustypeCar/truck/vehicle.Sell by-products of animals e.g. meat,
eggs, milk
.
## wherbusop_bustypeDoor to door/Go to customers.Sell by-products of animals
```

```
e.g. meat, eggs, milk
.
## wherbusop_bustypeFarm/small holding.Sell by-products of animals e.g. meat,
eggs, milk
.
## wherbusop_bustypeOffice block/office park.Sell by-products of animals e.g.
meat, eggs, milk
.
## wherbusop_bustypeOnline - internet, phone selling.Sell by-products of
animals e.g. meat, eggs, milk
.
## wherbusop_bustypeOpen space-Isipingo.Sell by-products of animals e.g.
meat, eggs, milk
.
## wherbusop_bustypeResidential premises - dwelling/garage/building on
residential premises.Sell by-products of animals e.g. meat, eggs, milk
.
## wherbusop_bustypeSchool Cafeteria.Sell by-products of animals e.g. meat,
eggs, milk
.
## wherbusop_bustypeShopping mall.Sell by-products of animals e.g. meat,
eggs, milk
.
## wherbusop_bustypeStall/table/container in a designated trading or market
area.Sell by-products of animals e.g. meat, eggs, milk
.
## wherbusop_bustypeStreet/street corner/pavement.Sell by-products of animals
e.g. meat, eggs, milk
.
## wherbusop_bustypeBusiness park/Premises dedicated to my business -
hotel/accommodation facility/factory/workshop.Sell something in the same form
that I buy from someone else (dont add value, e.g. cigarettes)  .
## wherbusop_bustypeCar/truck/vehicle.Sell something in the same form that I
buy from someone else (dont add value, e.g. cigarettes)
.
## wherbusop_bustypeDoor to door/Go to customers.Sell something in the same
form that I buy from someone else (dont add value, e.g. cigarettes)
.
## wherbusop_bustypeFarm/small holding.Sell something in the same form that I
buy from someone else (dont add value, e.g. cigarettes)
.
## wherbusop_bustypeOffice block/office park.Sell something in the same form
that I buy from someone else (dont add value, e.g. cigarettes)
.
## wherbusop_bustypeOnline - internet, phone selling.Sell something in the
same form that I buy from someone else (dont add value, e.g. cigarettes)
.
## wherbusop_bustypeOpen space-Isipingo.Sell something in the same form that
I buy from someone else (dont add value, e.g. cigarettes)
.
```

```
## wherbusop_bustypeResidential premises - dwelling/garage/building on
residential premises.Sell something in the same form that I buy from someone
else (dont add value, e.g. cigarettes)                                .
## wherbusop_bustypeSchool Cafeteria.Sell something in the same form that I
buy from someone else (dont add value, e.g. cigarettes)
.
## wherbusop_bustypeShopping mall.Sell something in the same form that I buy
from someone else (dont add value, e.g. cigarettes)
.
## wherbusop_bustypeStall/table/container in a designated trading or market
area.Sell something in the same form that I buy from someone else (dont add
value, e.g. cigarettes)                                              .
## wherbusop_bustypeStreet/street corner/pavement.Sell something in the same
form that I buy from someone else (dont add value, e.g. cigarettes)
.
## wherbusop_bustypeBusiness park/Premises dedicated to my business -
hotel/accommodation facility/factory/workshop.Sell something that I buy but
add value to, e.g. repackage, cook, etc                               .
## wherbusop_bustypeCar/truck/vehicle.Sell something that I buy but add value
to, e.g. repackage, cook, etc
.
## wherbusop_bustypeDoor to door/Go to customers.Sell something that I buy
but add value to, e.g. repackage, cook, etc
.
## wherbusop_bustypeFarm/small holding.Sell something that I buy but add
value to, e.g. repackage, cook, etc
.
## wherbusop_bustypeOffice block/office park.Sell something that I buy but
add value to, e.g. repackage, cook, etc
.
## wherbusop_bustypeOnline - internet, phone selling.Sell something that I
buy but add value to, e.g. repackage, cook, etc
.
## wherbusop_bustypeOpen space-Isipingo.Sell something that I buy but add
value to, e.g. repackage, cook, etc
.
## wherbusop_bustypeResidential premises - dwelling/garage/building on
residential premises.Sell something that I buy but add value to, e.g.
repackage, cook, etc                                                  .
## wherbusop_bustypeSchool Cafeteria.Sell something that I buy but add value
to, e.g. repackage, cook, etc
.
## wherbusop_bustypeShopping mall.Sell something that I buy but add value to,
e.g. repackage, cook, etc
.
## wherbusop_bustypeStall/table/container in a designated trading or market
area.Sell something that I buy but add value to, e.g. repackage, cook, etc
.
## wherbusop_bustypeStreet/street corner/pavement.Sell something that I buy
but add value to, e.g. repackage, cook, etc
```

```
.
## wherbusop_bustypeBusiness park/Premises dedicated to my business -
hotel/accommodation facility/factory/workshop.Sell something that I collect
from nature, e.g. herbs, firewood, charcoal, thatch, sand, stone  .
## wherbusop_bustypeCar/truck/vehicle.Sell something that I collect from
nature, e.g. herbs, firewood, charcoal, thatch, sand, stone
.
## wherbusop_bustypeDoor to door/Go to customers.Sell something that I
collect from nature, e.g. herbs, firewood, charcoal, thatch, sand, stone
.
## wherbusop_bustypeFarm/small holding.Sell something that I collect from
nature, e.g. herbs, firewood, charcoal, thatch, sand, stone
.
## wherbusop_bustypeOffice block/office park.Sell something that I collect
from nature, e.g. herbs, firewood, charcoal, thatch, sand, stone
.
## wherbusop_bustypeOnline - internet, phone selling.Sell something that I
collect from nature, e.g. herbs, firewood, charcoal, thatch, sand, stone
.
## wherbusop_bustypeOpen space-Isipingo.Sell something that I collect from
nature, e.g. herbs, firewood, charcoal, thatch, sand, stone
.
## wherbusop_bustypeResidential premises - dwelling/garage/building on
residential premises.Sell something that I collect from nature, e.g. herbs,
firewood, charcoal, thatch, sand, stone                              .
## wherbusop_bustypeSchool Cafeteria.Sell something that I collect from
nature, e.g. herbs, firewood, charcoal, thatch, sand, stone
.
## wherbusop_bustypeShopping mall.Sell something that I collect from nature,
e.g. herbs, firewood, charcoal, thatch, sand, stone
.
## wherbusop_bustypeStall/table/container in a designated trading or market
area.Sell something that I collect from nature, e.g. herbs, firewood,
charcoal, thatch, sand, stone                                        .
## wherbusop_bustypeStreet/street corner/pavement.Sell something that I
collect from nature, e.g. herbs, firewood, charcoal, thatch, sand, stone
.
## wherbusop_bustypeBusiness park/Premises dedicated to my business -
hotel/accommodation facility/factory/workshop.Sell something that I get for
free, e.g. second hand clothes, scrap metal                           .
## wherbusop_bustypeCar/truck/vehicle.Sell something that I get for free,
e.g. second hand clothes, scrap metal
.
## wherbusop_bustypeDoor to door/Go to customers.Sell something that I get
for free, e.g. second hand clothes, scrap metal
.
## wherbusop_bustypeFarm/small holding.Sell something that I get for free,
e.g. second hand clothes, scrap metal
.
## wherbusop_bustypeOffice block/office park.Sell something that I get for
```

free, e.g. second hand clothes, scrap metal
.
wherbusop_bustypeOnline - internet, phone selling.Sell something that I get for free, e.g. second hand clothes, scrap metal
.
wherbusop_bustypeOpen space-Isipingo.Sell something that I get for free, e.g. second hand clothes, scrap metal
.
wherbusop_bustypeResidential premises - dwelling/garage/building on residential premises.Sell something that I get for free, e.g. second hand clothes, scrap metal .
wherbusop_bustypeSchool Cafeteria.Sell something that I get for free, e.g. second hand clothes, scrap metal
.
wherbusop_bustypeShopping mall.Sell something that I get for free, e.g. second hand clothes, scrap metal
.
wherbusop_bustypeStall/table/container in a designated trading or market area.Sell something that I get for free, e.g. second hand clothes, scrap metal .
wherbusop_bustypeStreet/street corner/pavement.Sell something that I get for free, e.g. second hand clothes, scrap metal
.
wherbusop_bustypeBusiness park/Premises dedicated to my business - hotel/accommodation facility/factory/workshop.Sell something that I make e.g. crafts, clothes, furniture, bricks .
wherbusop_bustypeCar/truck/vehicle.Sell something that I make e.g. crafts, clothes, furniture, bricks
.
wherbusop_bustypeDoor to door/Go to customers.Sell something that I make e.g. crafts, clothes, furniture, bricks
.
wherbusop_bustypeFarm/small holding.Sell something that I make e.g. crafts, clothes, furniture, bricks
.
wherbusop_bustypeOffice block/office park.Sell something that I make e.g. crafts, clothes, furniture, bricks
.
wherbusop_bustypeOnline - internet, phone selling.Sell something that I make e.g. crafts, clothes, furniture, bricks
.
wherbusop_bustypeOpen space-Isipingo.Sell something that I make e.g. crafts, clothes, furniture, bricks
.
wherbusop_bustypeResidential premises - dwelling/garage/building on residential premises.Sell something that I make e.g. crafts, clothes, furniture, bricks .
wherbusop_bustypeSchool Cafeteria.Sell something that I make e.g. crafts, clothes, furniture, bricks
.

```
## wherbusop_bustypeShopping mall.Sell something that I make e.g. crafts,
clothes, furniture, bricks
.
## wherbusop_bustypeStall/table/container in a designated trading or market
area.Sell something that I make e.g. crafts, clothes, furniture, bricks
.
## wherbusop_bustypeStreet/street corner/pavement.Sell something that I make
e.g. crafts, clothes, furniture, bricks
.
## wherbusop_provCar/truck/vehicle.E.Cape
.
## wherbusop_provDoor to door/Go to customers.E.Cape
.
## wherbusop_provFarm/small holding.E.Cape
.
## wherbusop_provOffice block/office park.E.Cape
.
## wherbusop_provOnline - internet, phone selling.E.Cape
.
## wherbusop_provOpen space-Isipingo.E.Cape
.
## wherbusop_provResidential premises - dwelling/garage/building on
residential premises.E.Cape
-0.111106437
## wherbusop_provSchool Cafeteria.E.Cape
.
## wherbusop_provShopping mall.E.Cape
.
## wherbusop_provStall/table/container in a designated trading or market
area.E.Cape
.
## wherbusop_provStreet/street corner/pavement.E.Cape
.
## wherbusop_provBusiness park/Premises dedicated to my business -
hotel/accommodation facility/factory/workshop.Free State
.
## wherbusop_provCar/truck/vehicle.Free State
.
## wherbusop_provDoor to door/Go to customers.Free State
.
## wherbusop_provFarm/small holding.Free State
.
## wherbusop_provOffice block/office park.Free State
.
## wherbusop_provOnline - internet, phone selling.Free State
.
## wherbusop_provOpen space-Isipingo.Free State
.
## wherbusop_provResidential premises - dwelling/garage/building on
residential premises.Free State
```

```
.
## wherbusop_provSchool Cafeteria.Free State
.
## wherbusop_provShopping mall.Free State
.
## wherbusop_provStall/table/container in a designated trading or market
area.Free State
.
## wherbusop_provStreet/street corner/pavement.Free State
.
## wherbusop_provBusiness park/Premises dedicated to my business -
hotel/accommodation facility/factory/workshop.Gauteng
.
## wherbusop_provCar/truck/vehicle.Gauteng
.
## wherbusop_provDoor to door/Go to customers.Gauteng
.
## wherbusop_provFarm/small holding.Gauteng
.
## wherbusop_provOffice block/office park.Gauteng
.
## wherbusop_provOnline - internet, phone selling.Gauteng
.
## wherbusop_provOpen space-Isipingo.Gauteng
.
## wherbusop_provResidential premises - dwelling/garage/building on
residential premises.Gauteng
.
## wherbusop_provSchool Cafeteria.Gauteng
.
## wherbusop_provShopping mall.Gauteng
.
## wherbusop_provStall/table/container in a designated trading or market
area.Gauteng
.
## wherbusop_provStreet/street corner/pavement.Gauteng
.
## wherbusop_provBusiness park/Premises dedicated to my business -
hotel/accommodation facility/factory/workshop.KZN
0.005879545
## wherbusop_provCar/truck/vehicle.KZN
.
## wherbusop_provDoor to door/Go to customers.KZN
.
## wherbusop_provFarm/small holding.KZN
.
## wherbusop_provOffice block/office park.KZN
.
## wherbusop_provOnline - internet, phone selling.KZN
.
```

```
## wherbusop_provOpen space-Isipingo.KZN
.
## wherbusop_provResidential premises - dwelling/garage/building on
residential premises.KZN
.
## wherbusop_provSchool Cafeteria.KZN
.
## wherbusop_provShopping mall.KZN
.
## wherbusop_provStall/table/container in a designated trading or market
area.KZN
.
## wherbusop_provStreet/street corner/pavement.KZN
.
## wherbusop_provBusiness park/Premises dedicated to my business -
hotel/accommodation facility/factory/workshop.Limpopo
.
## wherbusop_provCar/truck/vehicle.Limpopo
.
## wherbusop_provDoor to door/Go to customers.Limpopo
.
## wherbusop_provFarm/small holding.Limpopo
.
## wherbusop_provOffice block/office park.Limpopo
.
## wherbusop_provOnline - internet, phone selling.Limpopo
.
## wherbusop_provOpen space-Isipingo.Limpopo
.
## wherbusop_provResidential premises - dwelling/garage/building on
residential premises.Limpopo
.
## wherbusop_provSchool Cafeteria.Limpopo
.
## wherbusop_provShopping mall.Limpopo
.
## wherbusop_provStall/table/container in a designated trading or market
area.Limpopo
.
## wherbusop_provStreet/street corner/pavement.Limpopo
.
## wherbusop_provBusiness park/Premises dedicated to my business -
hotel/accommodation facility/factory/workshop.Mpumalanga
.
## wherbusop_provCar/truck/vehicle.Mpumalanga
.
## wherbusop_provDoor to door/Go to customers.Mpumalanga
.
## wherbusop_provFarm/small holding.Mpumalanga
.
```

```
## wherbusop_provOffice block/office park.Mpumalanga
.
## wherbusop_provOnline - internet, phone selling.Mpumalanga
.
## wherbusop_provOpen space-Isipingo.Mpumalanga
.
## wherbusop_provResidential premises - dwelling/garage/building on
residential premises.Mpumalanga
.
## wherbusop_provSchool Cafeteria.Mpumalanga
.
## wherbusop_provShopping mall.Mpumalanga
.
## wherbusop_provStall/table/container in a designated trading or market
area.Mpumalanga
.
## wherbusop_provStreet/street corner/pavement.Mpumalanga
.
## wherbusop_provBusiness park/Premises dedicated to my business -
hotel/accommodation facility/factory/workshop.N.Cape
.
## wherbusop_provCar/truck/vehicle.N.Cape
.
## wherbusop_provDoor to door/Go to customers.N.Cape
.
## wherbusop_provFarm/small holding.N.Cape
.
## wherbusop_provOffice block/office park.N.Cape
.
## wherbusop_provOnline - internet, phone selling.N.Cape
.
## wherbusop_provOpen space-Isipingo.N.Cape
.
## wherbusop_provResidential premises - dwelling/garage/building on
residential premises.N.Cape
.
## wherbusop_provSchool Cafeteria.N.Cape
.
## wherbusop_provShopping mall.N.Cape
.
## wherbusop_provStall/table/container in a designated trading or market
area.N.Cape
.
## wherbusop_provStreet/street corner/pavement.N.Cape
.
## wherbusop_provBusiness park/Premises dedicated to my business -
hotel/accommodation facility/factory/workshop.N.West
.
## wherbusop_provCar/truck/vehicle.N.West
.
```

```
## wherbusop_provDoor to door/Go to customers.N.West
.
## wherbusop_provFarm/small holding.N.West
.
## wherbusop_provOffice block/office park.N.West
.
## wherbusop_provOnline - internet, phone selling.N.West
.
## wherbusop_provOpen space-Isipingo.N.West
.
## wherbusop_provResidential premises - dwelling/garage/building on
residential premises.N.West
.
## wherbusop_provSchool Cafeteria.N.West
.
## wherbusop_provShopping mall.N.West
.
## wherbusop_provStall/table/container in a designated trading or market
area.N.West
.
## wherbusop_provStreet/street corner/pavement.N.West
.
## wherbusop_provBusiness park/Premises dedicated to my business -
hotel/accommodation facility/factory/workshop.W.Cape
.
## wherbusop_provCar/truck/vehicle.W.Cape
.
## wherbusop_provDoor to door/Go to customers.W.Cape
.
## wherbusop_provFarm/small holding.W.Cape
.
## wherbusop_provOffice block/office park.W.Cape
.
## wherbusop_provOnline - internet, phone selling.W.Cape
.
## wherbusop_provOpen space-Isipingo.W.Cape
.
## wherbusop_provResidential premises - dwelling/garage/building on
residential premises.W.Cape
.
## wherbusop_provSchool Cafeteria.W.Cape
.
## wherbusop_provShopping mall.W.Cape
.
## wherbusop_provStall/table/container in a designated trading or market
area.W.Cape
.
## wherbusop_provStreet/street corner/pavement.W.Cape
.
## bustype_provFree State.Grow something and sell, e.g. fruit, vegetables,
```

```
plants (like a nursery)
.
## bustype_provGauteng.Grow something and sell, e.g. fruit, vegetables,
plants (like a nursery)
.
## bustype_provKZN.Grow something and sell, e.g. fruit, vegetables, plants
(like a nursery)
.
## bustype_provLimpopo.Grow something and sell, e.g. fruit, vegetables,
plants (like a nursery)
.
## bustype_provMpumalanga.Grow something and sell, e.g. fruit, vegetables,
plants (like a nursery)
.
## bustype_provN.Cape.Grow something and sell, e.g. fruit, vegetables, plants
(like a nursery)
.
## bustype_provN.West.Grow something and sell, e.g. fruit, vegetables, plants
(like a nursery)
.
## bustype_provW.Cape.Grow something and sell, e.g. fruit, vegetables, plants
(like a nursery)
.
## bustype_provE.Cape.Rear livestock/poultry and sell e.g. chickens
.
## bustype_provFree State.Rear livestock/poultry and sell e.g. chickens
.
## bustype_provGauteng.Rear livestock/poultry and sell e.g. chickens
.
## bustype_provKZN.Rear livestock/poultry and sell e.g. chickens
.
## bustype_provLimpopo.Rear livestock/poultry and sell e.g. chickens
.
## bustype_provMpumalanga.Rear livestock/poultry and sell e.g. chickens
.
## bustype_provN.Cape.Rear livestock/poultry and sell e.g. chickens
.
## bustype_provN.West.Rear livestock/poultry and sell e.g. chickens
.
## bustype_provW.Cape.Rear livestock/poultry and sell e.g. chickens
.
## bustype_provE.Cape.Render a professional service e.g. doctor, lawyer,
accountant, engineer, consultant
.
## bustype_provFree State.Render a professional service e.g. doctor, lawyer,
accountant, engineer, consultant
.
## bustype_provGauteng.Render a professional service e.g. doctor, lawyer,
accountant, engineer, consultant
.
```

```
## bustype_provKZN.Render a professional service e.g. doctor, lawyer,
accountant, engineer, consultant
.
## bustype_provLimpopo.Render a professional service e.g. doctor, lawyer,
accountant, engineer, consultant
.
## bustype_provMpumalanga.Render a professional service e.g. doctor, lawyer,
accountant, engineer, consultant
.
## bustype_provN.Cape.Render a professional service e.g. doctor, lawyer,
accountant, engineer, consultant
.
## bustype_provN.West.Render a professional service e.g. doctor, lawyer,
accountant, engineer, consultant
.
## bustype_provW.Cape.Render a professional service e.g. doctor, lawyer,
accountant, engineer, consultant
.
## bustype_provE.Cape.Render a skilled service e.g. mechanic, plumber, hair
salon, barber, painting, landscaping
.
## bustype_provFree State.Render a skilled service e.g. mechanic, plumber,
hair salon, barber, painting, landscaping
.
## bustype_provGauteng.Render a skilled service e.g. mechanic, plumber, hair
salon, barber, painting, landscaping
.
## bustype_provKZN.Render a skilled service e.g. mechanic, plumber, hair
salon, barber, painting, landscaping
.
## bustype_provLimpopo.Render a skilled service e.g. mechanic, plumber, hair
salon, barber, painting, landscaping
.
## bustype_provMpumalanga.Render a skilled service e.g. mechanic, plumber,
hair salon, barber, painting, landscaping
.
## bustype_provN.Cape.Render a skilled service e.g. mechanic, plumber, hair
salon, barber, painting, landscaping
.
## bustype_provN.West.Render a skilled service e.g. mechanic, plumber, hair
salon, barber, painting, landscaping
.
## bustype_provW.Cape.Render a skilled service e.g. mechanic, plumber, hair
salon, barber, painting, landscaping
.
## bustype_provE.Cape.Render building/construction services
.
## bustype_provFree State.Render building/construction services
.
## bustype_provGauteng.Render building/construction services
```

```
.
## bustype_provKZN.Render building/construction services
.
## bustype_provLimpopo.Render building/construction services
.
## bustype_provMpumalanga.Render building/construction services
.
## bustype_provN.Cape.Render building/construction services
.
## bustype_provN.West.Render building/construction services
.
## bustype_provW.Cape.Render building/construction services
.
## bustype_provE.Cape.Render other services e.g. car wash, garden services,
transport (taxi services), catering
.
## bustype_provFree State.Render other services e.g. car wash, garden
services, transport (taxi services), catering
.
## bustype_provGauteng.Render other services e.g. car wash, garden services,
transport (taxi services), catering
.
## bustype_provKZN.Render other services e.g. car wash, garden services,
transport (taxi services), catering
.
## bustype_provLimpopo.Render other services e.g. car wash, garden services,
transport (taxi services), catering
.
## bustype_provMpumalanga.Render other services e.g. car wash, garden
services, transport (taxi services), catering
.
## bustype_provN.Cape.Render other services e.g. car wash, garden services,
transport (taxi services), catering
.
## bustype_provN.West.Render other services e.g. car wash, garden services,
transport (taxi services), catering
.
## bustype_provW.Cape.Render other services e.g. car wash, garden services,
transport (taxi services), catering
.
## bustype_provE.Cape.Render tourism-related services e.g.
accommodation/hotel/B&B/guest house, tour operators
.
## bustype_provFree State.Render tourism-related services e.g.
accommodation/hotel/B&B/guest house, tour operators
.
## bustype_provGauteng.Render tourism-related services e.g.
accommodation/hotel/B&B/guest house, tour operators
.
## bustype_provKZN.Render tourism-related services e.g.
```

```
accommodation/hotel/B&B/guest house, tour operators
.
## bustype_provLimpopo.Render tourism-related services e.g.
accommodation/hotel/B&B/guest house, tour operators
.
## bustype_provMpumalanga.Render tourism-related services e.g.
accommodation/hotel/B&B/guest house, tour operators
.
## bustype_provN.Cape.Render tourism-related services e.g.
accommodation/hotel/B&B/guest house, tour operators
.
## bustype_provN.West.Render tourism-related services e.g.
accommodation/hotel/B&B/guest house, tour operators
.
## bustype_provW.Cape.Render tourism-related services e.g.
accommodation/hotel/B&B/guest house, tour operators
.
## bustype_provE.Cape.Sell by-products of animals e.g. meat, eggs, milk
.
## bustype_provFree State.Sell by-products of animals e.g. meat, eggs, milk
.
## bustype_provGauteng.Sell by-products of animals e.g. meat, eggs, milk
.
## bustype_provKZN.Sell by-products of animals e.g. meat, eggs, milk
.
## bustype_provLimpopo.Sell by-products of animals e.g. meat, eggs, milk
.
## bustype_provMpumalanga.Sell by-products of animals e.g. meat, eggs, milk
.
## bustype_provN.Cape.Sell by-products of animals e.g. meat, eggs, milk
.
## bustype_provN.West.Sell by-products of animals e.g. meat, eggs, milk
.
## bustype_provW.Cape.Sell by-products of animals e.g. meat, eggs, milk
.
## bustype_provE.Cape.Sell something in the same form that I buy from someone
else (dont add value, e.g. cigarettes)
.
## bustype_provFree State.Sell something in the same form that I buy from
someone else (dont add value, e.g. cigarettes)
.
## bustype_provGauteng.Sell something in the same form that I buy from
someone else (dont add value, e.g. cigarettes)
.
## bustype_provKZN.Sell something in the same form that I buy from someone
else (dont add value, e.g. cigarettes)
.
## bustype_provLimpopo.Sell something in the same form that I buy from
someone else (dont add value, e.g. cigarettes)
.
```

```
## bustype_provMpumalanga.Sell something in the same form that I buy from
someone else (dont add value, e.g. cigarettes)
.
## bustype_provN.Cape.Sell something in the same form that I buy from someone
else (dont add value, e.g. cigarettes)
.
## bustype_provN.West.Sell something in the same form that I buy from someone
else (dont add value, e.g. cigarettes)
.
## bustype_provW.Cape.Sell something in the same form that I buy from someone
else (dont add value, e.g. cigarettes)
.
## bustype_provE.Cape.Sell something that I buy but add value to, e.g.
repackage, cook, etc
.
## bustype_provFree State.Sell something that I buy but add value to, e.g.
repackage, cook, etc
.
## bustype_provGauteng.Sell something that I buy but add value to, e.g.
repackage, cook, etc
.
## bustype_provKZN.Sell something that I buy but add value to, e.g.
repackage, cook, etc
.
## bustype_provLimpopo.Sell something that I buy but add value to, e.g.
repackage, cook, etc
.
## bustype_provMpumalanga.Sell something that I buy but add value to, e.g.
repackage, cook, etc
.
## bustype_provN.Cape.Sell something that I buy but add value to, e.g.
repackage, cook, etc
.
## bustype_provN.West.Sell something that I buy but add value to, e.g.
repackage, cook, etc
.
## bustype_provW.Cape.Sell something that I buy but add value to, e.g.
repackage, cook, etc
.
## bustype_provE.Cape.Sell something that I collect from nature, e.g. herbs,
firewood, charcoal, thatch, sand, stone
.
## bustype_provFree State.Sell something that I collect from nature, e.g.
herbs, firewood, charcoal, thatch, sand, stone
.
## bustype_provGauteng.Sell something that I collect from nature, e.g. herbs,
firewood, charcoal, thatch, sand, stone
.
## bustype_provKZN.Sell something that I collect from nature, e.g. herbs,
firewood, charcoal, thatch, sand, stone
```

```
.
## bustype_provLimpopo.Sell something that I collect from nature, e.g. herbs,
firewood, charcoal, thatch, sand, stone
.
## bustype_provMpumalanga.Sell something that I collect from nature, e.g.
herbs, firewood, charcoal, thatch, sand, stone
.
## bustype_provN.Cape.Sell something that I collect from nature, e.g. herbs,
firewood, charcoal, thatch, sand, stone
.
## bustype_provN.West.Sell something that I collect from nature, e.g. herbs,
firewood, charcoal, thatch, sand, stone
.
## bustype_provW.Cape.Sell something that I collect from nature, e.g. herbs,
firewood, charcoal, thatch, sand, stone
.
## bustype_provE.Cape.Sell something that I get for free, e.g. second hand
clothes, scrap metal
.
## bustype_provFree State.Sell something that I get for free, e.g. second
hand clothes, scrap metal
.
## bustype_provGauteng.Sell something that I get for free, e.g. second hand
clothes, scrap metal
.
## bustype_provKZN.Sell something that I get for free, e.g. second hand
clothes, scrap metal
.
## bustype_provLimpopo.Sell something that I get for free, e.g. second hand
clothes, scrap metal
.
## bustype_provMpumalanga.Sell something that I get for free, e.g. second
hand clothes, scrap metal
.
## bustype_provN.Cape.Sell something that I get for free, e.g. second hand
clothes, scrap metal
.
## bustype_provN.West.Sell something that I get for free, e.g. second hand
clothes, scrap metal
.
## bustype_provW.Cape.Sell something that I get for free, e.g. second hand
clothes, scrap metal
.
## bustype_provE.Cape.Sell something that I make e.g. crafts, clothes,
furniture, bricks
.
## bustype_provFree State.Sell something that I make e.g. crafts, clothes,
furniture, bricks
.
## bustype_provGauteng.Sell something that I make e.g. crafts, clothes,
```

```
furniture, bricks
.
## bustype_provKZN.Sell something that I make e.g. crafts, clothes,
furniture, bricks
.
## bustype_provLimpopo.Sell something that I make e.g. crafts, clothes,
furniture, bricks
.
## bustype_provMpumalanga.Sell something that I make e.g. crafts, clothes,
furniture, bricks
.
## bustype_provN.Cape.Sell something that I make e.g. crafts, clothes,
furniture, bricks
.
## bustype_provN.West.Sell something that I make e.g. crafts, clothes,
furniture, bricks
.
## bustype_provW.Cape.Sell something that I make e.g. crafts, clothes,
furniture, bricks
.
```

Results and Discussion

Logged Regression model results discussion

Above, we have fitted two models, mainly the plain logged regression model, which was cross-validated to ensure that the model was not overfitted and there was no high variability in the model on unseen data. From the model, we restricted the variables to those that have a significance level of at least 0.05, which is the academic standard for the variables that will be accepted as having a significant relationship with profitability.

The first variable that has a significant relationship with business profitability is the variable pertaining to where a business is located, with rural, rural formal, and urban, both formal and informal. This variable shows that urban formal and informal have a positive relationship with profitability, though it was surprising that urban informal made the cut. This can be attributed to the fact that businesses located in urban areas have the advantage of being closer to customers. Furthermore, since urban areas tend to be clustered, they have a lot of customers within a short distance radius, relative to their non-urban counterparts, thus they have a relatively bigger client base to supply goods to or services. Additionally, also related to location, businesses that operate in open spaces also seemed to have a positive relationship with business performance, however upon closer inspection, this category has only one value, which is not enough to determine significance.

This analysis further looked into another variable, which pertains to the total amount of hours worked. The variable has a positive relationship with business performance, the more hours worked, the higher the profitability of the business though the value is relatively small at 0.015888976, meaning a unit increase in total hours worked is associated with just 1.59 % increase in profitability, which is comparatively lower than the other factors. Of course, no factor is small as when combined with others can drive a lot of

growth in profitability. Thus, it seems working longer hours is positively associated with growth, though this could be a challenge in townships where crime is an issue, especially if you operate too early in the morning or too late.

Moving forward, we look into the variables that pertain to whom the business sells to, with selling to individuals, selling to other small businesses, larger businesses, and the government as the various columns we have, and a binary choice between selling or not selling. Looking into the first variable in this group of variables, we see there's a very significant negative relationship between selling to private individuals and business performance, with not selling to private individuals associated with a 56.59% increase in profitability. This implies that SMMEs tend to struggle with competing with the core economy comprised of larger businesses and the government when it comes to servicing the population. However, this is not surprising in South Africa since there's a relatively developed core economy. Further, we also see a major significant positive relationship between selling to larger businesses and the government, with an instance of selling to these entities respectively associated with a 26.37% and 63.87% increase in profitability. This shows that SMMEs are much more profitable when they use these bigger businesses as sources of cash and provide them with services or goods that the bigger businesses and government can't in-source and they are better off sourcing them independently. The next variable, which looks at whether a business's tender bid was successful or not shows that there's a negative relationship between having an unsuccessful tender application and business performance. This feeds on the observations above which shows that SMMEs that service bigger businesses and the government tend to be successful. Thus, having a tender unsuccessful will tend to have devastating impacts on the business's profits.

Another variable of interest looked into the education cluster of variables, which shed light on managerial competence. The general consensus among all the variables is that education has a positive relationship with business performance. Further, we see that both in terms of the p-value and the co-efficient having at least an education(Some Primary education) and level 6 (Post matric qualification) which is the second highest level of education have the greatest and most robust effect on the performance of businesses in South Africa. Meaning, that while on the one hand, being literate can have a huge impact on the performance of businesses relative to the illiterate counterparts, having at least a post-matric qualification also has a positive impact on performance. The post-matric qualification factor points out that skills play a pivotal role in the value each enterprise offers and thus its profitability.

We further shift our attention to the access to financial services variable, which looks at whether a business is served or not in the financial services by entities such as banks, microfinance institutions, and many more. This variable also asses whether those services are formal or informal. Overall, only one of the categories has a significant negative relationship with business performance, and that category is *not being served*. This makes sense as financial access gives a boost to businesses in their ability to execute their business objectives. Services such as banking, having a credit line, insurance products, and if they have employees, services such as payroll systems are valuable to businesses and can improve the constraints to trading and enable ease of doing business. Access to financial services plays a critical role in ensuring overall that the business is able to grow, compete

for contracts, and invest in their business either through borrowing or seeking out investments.

Another variable that relates to access to finance looks into those businesses that either borrow from their friends or family as a source of financing, implying that these businesses don't have adequate access to proper sources of finance to aid in their growth. This variable shows that businesses that rely on borrowing from friends and family have a negative relationship with business performance, with such businesses associated with a 39.62% decline in profitability. This further shines a light on the importance of having adequate sources of financing to bolster growth. This could be the case as friends and families are not financial institutions, they will tend to be cash-constrained and thus do not have the adequate financial muscle to help finance an entire business operation.

We further see the variable that looks into whether a business has computerized financial records or not is also a significant indicator of profitability. As we explained above, keeping records should have a robust relationship with performance as such businesses have a trail of their operations, assets, and past financial transactions which can help the business keep track of past performances and financial information. This information can be of value to financiers and creditors who can utilize such information to assess the investability of businesses. Further, businesses can use such information to plan and optimize their performance by looking at where they perform best and focusing on those areas that can further drive growth going forward.

The next variable looks into whether a person owns their private residence or not. Surprisingly, this variable has a negative relationship with business performance, which is surprising as we'd expect owners that have houses should be able to pledge them as collateral, and thus be able to access credit markets and invest such monies to their business.

Another significant variable has to do with enterprise classification, showing that micro-enterprises that tend to have at least one employee tend to have a positive relationship with business performance. This makes sense as this category when we ran the analysis, had a fair amount of data compared to other categories such as the small and medium categories and thus the model can pick the variability it brings. But, also, it makes sense that this is the case as micro-enterprises will be comprised of businesses that unlike their own account counterparts who also contained a lot of data points in ours, were inspired by taking advantage of business opportunities rather than going into businesses for survival ends. Thus, these enterprises would be relatively profitable compared to their own account businesses. Micro enterprises will tend to be your spaza shops, wholesalers, bottle stores, funeral parlors, mini construction businesses, and many other like businesses. Owners of these businesses would have invested a fairly large amount of money into them, and they would be responding to local opportunities.

Next, we look into the age variable, which assesses how long an enterprise has been in operation. The variable shows a positive relationship between performance and the age of a business, though not very strong as a unit increase in the age of an enterprise is associated with a 0.8561% increase in their profitability. This positive relationship can be

attributed to the experience that each business has in a particular industry and how efficient they are in pursuing the opportunities in that industry due to their accumulated experience in that particular industry. Further, if a business survives the first initial years, and builds brand awareness, then that business is on its way to higher performance than those counterparts that fail in the earlier years of operations.

The next set of variables we deal with that are significant has to do with interactions between variables. Interactive variables are the variables that move together, and their significance is contingent on their relationship with each other. The first interaction looks into the relationship between where the business operates and the services it renders. This interaction significantly shows a negative relationship between those businesses that sell in stalls, containers, or market areas and offer other services that are not highly skilled as having a relatively weaker relationship compared to our baseline. Thus, these businesses tend to be not powerful sources of profitability.

The next interaction we look into has to do with relationships between businesses that operate in areas dedicated to business operations such as a hotel, accommodation facilities, or a factory and selling by-products of animals. These are your butcheries, egg sellers, poultry sellers, and many other animal by-products. This variable shows a positive relationship between business performance and this interaction, showing the importance of having dedicated business premises and selling things that require a relatively high level of skill.

Another interaction of interest pertains to the instance where a business is located on a farm or stall and interacts with buying stuff and selling it, such as cooking. This interaction and business performance is negative, and it makes sense since there's no readily available market where the farms are located for these products. The next set of interaction terms has to do with where a business operates and the province it is located in. Specifically, we see a positive interaction between businesses that operate in stalls or farms and are located in the following provinces, Eastern Cape, KZN, Limpopo, and Gauteng with business performance. This is logical as these provinces have extensive farmland that can be used to generate value and sell, and thus generate profits and thus farming in these provinces tend to thrive.

Lasso Regression

From the lasso model that we fitted, its main task was to ensure that we did not overfit the data and we further cross-validated to ensure that we were not biased toward one training set and that our results could generalize by minimizing errors across different test sets, and also pick the lambda parameter from the cross-validated errors. The model forced all the less significant variables to zero and thus a few variables were chosen for the final model.

The first variable it picked was the formal urban variable, and this variable has a positive relationship with business performance. Furthermore, this variable was also significant in the logged model. Thus, it is clear that businesses that are located in the urban and formal areas perform relatively better than their rural, rural-formal, and urban informal counterparts. This as we outlined in the variable construction part of this analysis, can be attributed to the relatively more developed infrastructure in the formal urban areas which

can support businesses in their operations, such as better roads, business premises or industrial zones, and supply chains that are nearby which can help these businesses optimize operations and perform better. Further, and most importantly, urban areas have readily available networks of other businesses and independent consumers, these businesses can supply to and or offer services to.

Further, moving forward, our analysis looks into the variable that probes where a business is located. Though, however, this variable is not significant across both models, it is still worth looking into. This variable points out that, in the instances where the business operates in residential premises as having a negative relationship with business performance, such occurrence is associated with a 3.55% decrease in profitability. This is an indication that businesses that have just started should probably look for more business-appropriate premises, especially if they are an SMME where there's possibly a high influx of clients to buy their goods and services and residential premises are not appropriate to run a business.

Our analysis also notes the set of variables that pertains to education which shows managerial competence. These variables point out that having at least some primary education is robust, even in the lasso regression case of business performance. This variable was also significant in the case of the logged regression, but it was so across different categories not just the first level. This solidifies the importance of at least having some literacy in running a business.

Another variable that is coming as significant, and was also forming part of the variables we picked from the baseline model has to do with not having access to financial services as having a negative relationship with business performance.

Furthermore, the variable that probes whether a business is registered or not is also significant for the lasso model. We used the variable to signal whether a business is formal or not when we defined our variables. Though this variable is not collaborated by both models, this result shows that businesses that are formal will be positively related to performance compared to their unregistered counterparts. This has to do with the kind of opportunities that such businesses have or rather can pursue, such as access to banking services, the ability to apply for contracts and tenders, and many other advantages of having a registered enterprise. Thus, they are able to capitalize on these and be relatively more successful. Further, it could be argued that businesses that are registered are relatively more formal, thus compared to their more informal counterparts, they are rather motivated by profiteering from opportunities from the markets rather than survival imperatives.

Next, this analysis moves towards the variable that looks at whether a business keeps computerized financial records or not. This variable is significant across both models and it has to do with the quality of the records the business keeps. As explained above, we expect that a business that has quality records, which are computerized and thus are structured will probably be of high value to investors and potential creditors. Thus, as explained above, drives business growth and ultimately business performance. Further, together with education, we expect businesses that have managers who take the initiative to seek out

expertise in data management systems to be competent managers, and thus will all round better manage the business.

Another variable, specifically from the lasso model but not collaborated by both models is the variable that looks into the business access and usage of insurance products. The first variable pertaining to insurance looks into whether a business has an insurance product, having insurance associated with the presence of an insurance product associated with a 10.3% increase in the profitability of a business, and a unit increase in the number of insurance products each business has also associated with a 0.255% increase in the profitability of businesses. The importance of insurance pertains to the risk management attitudes of businesses, whether or not they have mechanisms in place to deal with the occurrence of adverse events such as theft, vandalism, or any crime in general. Further, such measures can also be enticing to potential investors as such measures signal that the owners of such businesses care about losses that might be incurred to such an extent that they are taking proactive measures, which would protect the stake of the investors. Further, depending on the type taken, insurance can place a floor on the amount of losses that can be incurred by protecting against unforeseen events.

Now, we look at another variable that is related to the above-mentioned related to having measures against crime. The variable is positively associated with business performance, with the presence of such measures, though it does not specify which ones, are associated with a staggering 16.47% increase in the profits of SMMEs. The causal link between measures against crime and profitability is twofold. One has to do with the fact that businesses that have such measures in the first place might be already successful and thus have to protect their assets which protect their assets against crime. Or two, businesses that prioritize measures against crime become successful because they tend to be able to retain their assets for a long period of time, and thus be able to deliver long-term growth. Furthermore, businesses that are able to protect their assets will tend to be able to attract capital and debt as their risk profile is relatively less compared to those who don't have adequate risk management procedures.

Though insurance and measures against crime are important, another variable that also has a significant impact on performance, and has to do with the quality of the business is the *access to how many business functions variable*. This variable looks into how many of the important business functions a business has access to, such as a vision and mission statement, business plan, business strategy, marketing plan, accounting systems, formal training of staff, and business budgeting. From the model, we see that a unit increase in the number of these functions is associated with a 0.712% increase in the profitability of a business. This shows that a business that has at least a plan on how to take advantage of opportunities in the market documented will tend to perform better than those who don't have such measures. Such effort signals that a business has taken the time to recognize opportunities that exist in the market, and figured out a way to go about executing them. These are also a testament to the managerial competence in taking advantage of market opportunities by carefully planning how to go about and exploit them.

Another variable that has a positive relationship with business performance pertains to the enterprise classification. This variable shows that there's a significant relationship between

being at the least a micro-enterprise, which is just a step above the own account category, with this variable showing that being at the least a micro-enterprise is associated with a 64.12% increase in profitability. This insight collaborates with the result we gained from the logged model above which also highlighted the same relationship. The insight still remains here, that micro-enterprises unlike their own account counterparts are motivated by taking advantage of opportunities in the market rather than being in business for survival reasons since the owners can't get opportunities in the market.

The last set of variables that we look into from the lasso regression model pertains to the interaction variables, where the importance of each variable is contingent on another variable and vice versa. We see for instance business premises as being important when they interact with a province. For instance, we see that having your business hosted in your residential premises and being in the EC are negatively correlated, with such occurrences associated with an 11.11% decline in profitability. Also, we also see that KZN and businesses that operate in residential premises dedicated to a business such as a factory, business park, or hotel have a significant relationship with profitability.

Conclusion

Inspired by the issue of chronic unemployment in South Africa, especially among the youth, this analysis seeks to analyze the factors that are important in driving SMME performance in South Africa, using business profitability as the proxy for business performance. We looked at SMMEs as they have the greatest potential for driving growth and improving the employment landscape and they further form an integral part of government policy to tackle unemployment.

One of the core findings of this study across both models we utilized, the lasso and the logged model is that SMMEs that sell goods or services to larger businesses and the government tend to perform relatively better compared to their counterparts who don't. Further, those enterprises that offer services to private individuals tend to have a negative relationship with business performance. Further, location, specifically being located in the urban formal sector is also a significant predictor of business performance, and this was also collaborated by both models. This from the analysis can be due to the strategic location that urban areas find themselves in, both in terms of infrastructure and density of clients. Further, related to location, our models show that it is important to host your business in business-appropriate premises and in line with your business such as in a factory, business park, or hotels or BnB for accommodation businesses, avoiding residential areas as they negatively linked with business performance..

Another strong theme in this analysis we came across was the theme of managerial competence. We saw that across both models being at the least literate is associated with more profitability. Other proxies for competence relate to the strategic decisions that business owners take on behalf of their businesses, such as taking out insurance products, having measures against crime, accessing business functions such as accounting systems, budgeting, and many others. All these proxies were found to be significantly associated with profitability. Furthermore, it's clear that a culmination of these variables is important in having a profitable enterprise.

Our analysis further showed that, and this insight was collaborated by both models on the importance of being at least a micro-enterprise as having a positive effect on the profitability of firms. This can be attributed that in the context of South Africa where there is high unemployment people who start own account enterprises are mostly motivated by survival imperatives and would opt for a job in the formal sector given the chance. Thus, on the other hand, micro-enterprises are more likely to be motivated by chasing market opportunities rather than survival imperatives. Thus, taking time to investigate the market you want to enter and investing a considerable amount of finances to make your business more appropriate for the market can go a long way.

Recommendations

It should be noted though that this is a relatively old dataset and thus some of the variables are outdated and reflect the state of the world when the data was collected. Thus, new data might have come to the fore with a newer set of variables that are more robust and reflective of the contemporary period. Thus, this study should be repeated with much newer datasets that reflect the status quo, i.e. the advent of smartphones, social media, financial technology advancements, and many more that are relevant to businesses that operate in the 2020s.