Name& Surname : Mandla Hadebe

Name of Project : Insurance

#######################################

#Swedish Motor Insurance Data Analysis

#######################################

install.packages("ggplot2")

library("ggplot2") #For data visualizations

install.packages("plyr")

library(plyr)

install.packages("dplyr")

library(dplyr)

#Attribute #Description

#Kilometres kilometres travelled per year

#Zones Geographical zone

#Bonus No claims bonus; equal to the number of years, plus one, since the last claim

#Make 1-8 represents eight different common car models. All other models are combined in class 9

#Insured Number of insured in policy-years

#Claims Number of claims

#Payment Total value of payments in Skr (Swedish Krona)

########################

#1# Loading Data and Analysis

########################

#getwd

insurance= read.csv("SwedishMotorInsurance.csv",header=T, sep=',')

#insurance= read.csv("SwedishMotorInsurance.csv",header=T, sep=',')

head(insurance)

#Kilometres Zone Bonus Make Insured Claims Payment

#1 1 1 1 1 455.13 108 392491

#2 1 1 1 2 69.17 19 46221

#3 1 1 1 3 72.88 13 15694

#4 1 1 1 4 1292.39 124 422201

#5 1 1 1 5 191.01 40 119373

#6 1 1 1 6 477.66 57 170913

nrow(insurance)

#[1] 2182

dim(insurance)

#[1] 2182 7 #Dimensionality of the dataset 2182 rows and 7 columns

summary(insurance)

#Kilometres Zone Bonus Make Insured

#Min. :1.000 Min. :1.00 Min. :1.000 Min. :1.000 Min. : 0.01

#1st Qu.:2.000 1st Qu.:2.00 1st Qu.:2.000 1st Qu.:3.000 1st Qu.: 21.61

#Median :3.000 Median :4.00 Median :4.000 Median :5.000 Median : 81.53

#Mean :2.986 Mean :3.97 Mean :4.015 Mean :4.992 Mean : 1092.20

#3rd Qu.:4.000 3rd Qu.:6.00 3rd Qu.:6.000 3rd Qu.:7.000 3rd Qu.: 389.78

#Max. :5.000 Max. :7.00 Max. :7.000 Max. :9.000 Max. :127687.27

#Claims Payment

#Min. : 0.00 Min. : 0

#1st Qu.: 1.00 1st Qu.: 2989

#Median : 5.00 Median : 27404

#Mean : 51.87 Mean : 257008

#3rd Qu.: 21.00 3rd Qu.: 111954

#Max. :3338.00 Max. :18245026

colnames(insurance) #Calling all the columns present in the data set

#[1] "Kilometres" "Zone" "Bonus" "Make" "Insured" "Claims" "Payment"

str(insurance) #To get the structure of the data

#'data.frame': 2182 obs. of 7 variables:

#$ Kilometres: int 1 1 1 1 1 1 1 1 1 1 ...

#$ Zone : int 1 1 1 1 1 1 1 1 1 1 ...

#$ Bonus : int 1 1 1 1 1 1 1 1 1 2 ...

#$ Make : int 1 2 3 4 5 6 7 8 9 1 ...

#$ Insured : num 455.1 69.2 72.9 1292.4 191 ...

#$ Claims : int 108 19 13 124 40 57 23 14 1704 45 ...

#$ Payment : int 392491 46221 15694 422201 119373 170913 56940 77487 6805992 214011 ...

# Kilometres, Zone, Bonus and Make are be categorical variable

#Converting (Kilometres, Zone, Bonus and Make) to factor variables since I'm working with categorical variable

insurance$Kilometres=as.factor(insurance$Kilometres)

insurance$Zone=as.factor(insurance$Zone)

insurance$Make=as.factor(insurance$Make)

insurance$Bonus = as.factor(insurance$Bonus)

summary(insurance) #Describe my entire data set

str(insurance) #Structure of dataset after converting categorical variable to factor

#data.frame': 2182 obs. of 7 variables:

#$ Kilometres: Factor w/ 5 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...

#$ Zone : Factor w/ 7 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...

#$ Bonus : Factor w/ 7 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 2 ...

#$ Make : Factor w/ 9 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 1 ...

#$ Insured : num 455.1 69.2 72.9 1292.4 191 ...

#$ Claims : int 108 19 13 124 40 57 23 14 1704 45 ...

#$ Payment : int 392491 46221 15694 422201 119373 170913 56940 77487 6805992 214011 ...

#summary(scale(insurance$Insured))

head(insurance) #Calling the first six rows

#Kilometres Zone Bonus Make Insured Claims Payment

#1 1 1 1 1 455.13 108 392491

#2 1 1 1 2 69.17 19 46221

#3 1 1 1 3 72.88 13 15694

#4 1 1 1 4 1292.39 124 422201

#5 1 1 1 5 191.01 40 119373

#6 1 1 1 6 477.66 57 170913

############################

#Looking for Missing Values

############################

a =c(1,2,3,4,NA,5)

is.na(a)

is.na(insurance)

head(is.na(insurance))

colSums(is.na(insurance)) #Columns with missing values

sum(is.na(insurance)) #sum of missing values

####################################################

#The total value of payment by an insurance company

####################################################

sum(insurance$Payment) #Sum of all the payments

sum(insurance$Claims) #Sum of all the claims

sum(insurance$Insured) #Sum of all the insured

#################################################################################

#2# payment is related to number of claims and the number of insured policy years

#################################################################################

insurance=read.csv("SwedishMotorInsurance.csv")

insurance

head(insurance)

plot(insurance$Claims,insurance$Payment) #Dataset visualization for Claims vs Payment

cor(insurance$Claims,insurance$Payment) #It tells me there is a strong positive correlation which mean increase in claims there is increase in payment value

#[1] 0.9954003

plot(insurance$Insured,insurance$Payment) #Dataset visualization for Claims vs Payment

cor(insurance$Insured,insurance$Payment) #Higher correlation tells me as the insured increase the payment value also increase

par(mfrow=c(1,2))

###Testing for linear relationship between payment and number of claims

#H0 :There is a no linear relationship between payment and number of claims

#H1 :There is a linear relationship between payment and number of claims

alpha =0.05

pvalue=2.2e-16

pvalue<alpha #Reject H0 if pvalue<alpha value

model=lm(Payment~Claims, data = insurance) #dependent~independent

summary(model)

#Call:

# lm(formula = Payment ~ Claims, data = insurance)

#Residuals:

# Min 1Q Median 3Q Max

#-1744858 -8545 2773 13386 1491369

#Coefficients:

# Estimate Std. Error t value Pr(>|t|)

#(Intercept) -3362.29 2154.79 -1.56 0.119

#Claims 5020.08 10.35 485.11 <2e-16 \*\*\*

# ---

# Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

#Residual standard error: 97480 on 2180 degrees of freedom

#Multiple R-squared: 0.9908, Adjusted R-squared: 0.9908 #R values show the accuracy of the model

#F-statistic: 2.353e+05 on 1 and 2180 DF, p-value: < 2.2e-16

pvalue<alpha

#Results: We reject H0,there is a relationship between payment and the number of claims

###Testing for linear relationship between payment and insured policy years

#H0 :There is a no linear relationship between payment and insured policy years

#H1 :There is a linear relationship between payment and insured policy years

alpha =0.05

pvalue=2.2e-16

pvalue<alpha #Reject H0 if pvalue<alpha value

model1=lm(Payment~Insured, data = insurance) #dependent~independent

summary(model1)

#Call:

# lm(formula = Payment ~ Insured, data = insurance)

#Residuals:

# Min 1Q Median 3Q Max

#-5946157 -75828 -70260 -30246 5343552

#Coefficients:

# Estimate Std. Error t value Pr(>|t|)

#(Intercept) 73852.388 7971.250 9.265 <2e-16 \*\*\*

# Insured 167.695 1.383 121.266 <2e-16 \*\*\*

# ---

# Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

#Residual standard error: 365600 on 2180 degrees of freedom

#Multiple R-squared: 0.8709, Adjusted R-squared: 0.8708 Rsqrd value shows that my model is generating correct values by 87%

#F-statistic: 1.471e+04 on 1 and 2180 DF, p-value: < 2.2e-16

pvalue<alpha #Reject H0 if pvalue <alpha

#Results: We reject H0,there is a relationship between payment and insured policy years

##########################################################################

#3# Findind ing whether location ,distance,make ,insured amount ,claims or bonus affect payment increase and decrease

###########################################################################

#H0:payment increase and decrease is not affected by location ,distance,make ,insured amount ,claims or bonus

#H1:payment increase and decrease is affected by location ,distance,make ,insured amount ,claims or bonus

alpha=0.05

pvalue<alpha #If factors are less than pvalue it tells us it affect the payment increase and decrease

model2=lm(Payment~Kilometres+Insured+Zone+Make+Insured+Claims+Bonus, data = insurance)##fix this

summary(model2)

#Where pvalue< alpha that factor affect the payment

#Call:

# lm(formula = Payment ~ Kilometres + Insured + Zone + Make + Insured +

# Claims + Bonus, data = insurance)

#Residuals:

# Min 1Q Median 3Q Max

#-762236 -18278 -1588 16179 831273

#Coefficients:

# Estimate Std. Error t value Pr(>|t|)

#(Intercept) -1.346e+04 7.596e+03 -1.772 0.076604 .

#Kilometres2 2.236e+04 4.707e+03 4.751 2.16e-06 \*\*\*

# Kilometres3 2.329e+04 4.703e+03 4.952 7.92e-07 \*\*\*

# Kilometres4 2.161e+04 4.746e+03 4.553 5.59e-06 \*\*\*

#Kilometres5 2.150e+04 4.770e+03 4.507 6.92e-06 \*\*\*

#Insured 2.809e+01 6.817e-01 41.206 < 2e-16 \*\*\*

#Zone2 1.402e+03 5.557e+03 0.252 0.800802

#Zone3 3.908e+03 5.568e+03 0.702 0.482789

#Zone4 3.314e+04 5.591e+03 5.927 3.58e-09 \*\*\*

# Zone5 6.512e+03 5.614e+03 1.160 0.246247

#Zone6 1.936e+04 5.598e+03 3.458 0.000555 \*\*\*

# Zone7 4.971e+03 5.740e+03 0.866 0.386572

#Make2 -1.527e+04 6.306e+03 -2.422 0.015521 \*

# Make3 -1.283e+04 6.330e+03 -2.026 0.042851 \*

# Make4 -2.647e+04 6.359e+03 -4.162 3.28e-05 \*\*\*

#Make5 -1.814e+04 6.312e+03 -2.873 0.004100 \*\*

#Make6 -1.863e+04 6.311e+03 -2.953 0.003185 \*\*

#Make7 -2.016e+04 6.329e+03 -3.186 0.001463 \*\*

#Make8 -1.005e+04 6.368e+03 -1.578 0.114664

#Make9 -6.808e+03 7.004e+03 -0.972 0.331169

#Claims 4.289e+03 2.089e+01 205.288 < 2e-16 \*\*\*

# Bonus2 3.880e+03 5.626e+03 0.690 0.490513

#Bonus3 4.371e+03 5.653e+03 0.773 0.439561

#Bonus4 1.063e+03 5.663e+03 0.188 0.851057

#Bonus5 -1.354e+03 5.648e+03 -0.240 0.810524

#Bonus6 3.863e+03 5.624e+03 0.687 0.492220

#Bonus7 1.536e+04 5.751e+03 2.670 0.007638 \*\*

# ---

# Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

#Residual standard error: 69670 on 2155 degrees of freedom

#Multiple R-squared: 0.9954, Adjusted R-squared: 0.9953 Rsqrd value tells me how accurate is my model

#F-statistic: 1.78e+04 on 26 and 2155 DF, p-value: < 2.2e-16

#Results: Kilometres affect the payment

#Insured affect the payment

#Zones does not affect the payment

#make affect the payment

#Claims affect the payment

##########################################################################

#4# insurance company is planning to establish a new branch office

##########################################################################

#Kilometers travelled per year 1: < 1000

#2: 1000-15000

#3: 15000-20000

#4:20000-25000

#5: > 25000

group1<-apply(insurance[,c(5,6,7)], 2, function(x) tapply(x, insurance$Zone, mean))

group1

group2<-apply(insurance[,c(5,6,7)], 2, function(x) tapply(x, insurance$Kilometres, mean))

group2

group3<-apply(insurance[,c(5,6,7)], 2, function(x) tapply(x, insurance$Bonus, mean))

group3

str(insurance)

# Result:

# The following observations can be made from the results:

# a. Zone 4 has the highest number of claims, and thus payment as well.

# b. Zones 1-4 have more insured years, claims, and payments.

# c. Kilometer group 2 has the maximum payments. Though the insured number

# of years is lesser than kilometre 1, the claims and payments are higher for group 2.

# d. There is not much variation in groups of bonus except for 7 with unusually

# high number of insured years, claims, and payments.

#################################################################################

#5# payment is related to number of claims and the number of insured policy years

#################################################################################

############################

#Linear Regresiion Analysis

###########################

insurance=read.csv("SwedishMotorInsurance.csv")

View(insurance)

summary(insurance)

str(insurance)

dim(insurance)

###Testing for linear relationship between Claims and insured amount, zone, kilometer, bonus, or make

#H0 :There is a no linear relationship between Claims and insured amount, zone, kilometer, bonus, or make

#H1 :There is a linear relationship between Claims and insured amount, zone, kilometer, bonus, or make

alpha =0.05

pvalue=2.2e-16

pvalue<alpha #Reject H0 if pvalue<alpha value

model3=lm(Claims~Insured+Zone+Kilometres+Bonus+Make, data = insurance)

summary(model3)

#pvalue<alpha If pvalue<pvalue of factors it tell us that it affacet the number of Claims

#Call:

# lm(formula = Claims ~ Insured + Zone + Kilometres + Bonus + Make, data = insurance)

#Residuals:

# Min 1Q Median 3Q Max

#-983.95 -16.36 0.06 14.09 1222.44

#Coefficients:

# Estimate Std. Error t value Pr(>|t|)

#(Intercept) 7.130e+01 7.679e+00 9.284 < 2e-16 \*\*\*

# Insured 2.924e-02 3.122e-04 93.649 < 2e-16 \*\*\*

# Zone2 -1.165e+01 5.724e+00 -2.036 0.041887 \*

#Zone3 -1.983e+01 5.724e+00 -3.464 0.000543 \*\*\*

#Zone4 -2.059e+01 5.747e+00 -3.583 0.000347 \*\*\*

#Zone5 -3.574e+01 5.737e+00 -6.230 5.60e-10 \*\*\*

#Zone6 -3.416e+01 5.724e+00 -5.969 2.79e-09 \*\*\*

#Zone7 -4.461e+01 5.839e+00 -7.641 3.23e-14 \*\*\*

#Kilometres2 1.423e+01 4.843e+00 2.938 0.003341 \*\*

#Kilometres3 8.060e-01 4.848e+00 0.166 0.867982

#Kilometres4 -1.317e+01 4.884e+00 -2.697 0.007057 \*\*

# Kilometres5 -1.309e+01 4.910e+00 -2.666 0.007737 \*\*

# Bonus2 -2.533e+01 5.775e+00 -4.385 1.21e-05 \*\*\*

#Bonus3 -3.334e+01 5.784e+00 -5.765 9.35e-09 \*\*\*

#Bonus4 -3.679e+01 5.784e+00 -6.361 2.44e-10 \*\*\*

#Bonus5 -3.614e+01 5.771e+00 -6.263 4.55e-10 \*\*\*

#Bonus6 -2.950e+01 5.763e+00 -5.119 3.35e-07 \*\*\*

#Bonus7 -2.374e+01 5.907e+00 -4.019 6.03e-05 \*\*\*

#Make2 -1.375e+01 6.494e+00 -2.117 0.034346 \*

#Make3 -1.727e+01 6.515e+00 -2.651 0.008088 \*\*

#Make4 -1.911e+01 6.543e+00 -2.921 0.003523 \*\*

#Make5 -1.278e+01 6.501e+00 -1.966 0.049478 \*

#Make6 -1.514e+01 6.498e+00 -2.330 0.019899 \*

#Make7 -1.611e+01 6.515e+00 -2.473 0.013469 \*

#Make8 -1.813e+01 6.553e+00 -2.767 0.005712 \*\*

#Make9 1.180e+02 6.759e+00 17.451 < 2e-16 \*\*\*

#---

#Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

#Residual standard error: 71.83 on 2156 degrees of freedom

#Multiple R-squared: 0.8746, Adjusted R-squared: 0.8732

#F-statistic: 601.7 on 25 and 2156 DF, p-value: < 2.2e-16

#Results: Insured affect the number of claims

#Zone affect the number of claims

#Some locations (not every location please note) affect the number of claims

#Bonus affect the number of claims