#####IS 470 Data

Exploration---------------------------------------------------------

# Part 1: Pokemon Data ###

# This dataset contains information on 800 Pokemon from six generations of Pokemon.

# VARIABLE DESCRIPTIONS:

#number: The entry number of the Pokemon #name: The English name of the Pokemon #typel: The Primary Type of the Pokemon #type2: The Secondary Type of the Pokemon #hp: The Base HP of the Pokemon

#attack: The Base Attack of the Pokemon #defense: The Base Defense of the Pokemon #sp.atk: The Base Special Attack of the Pokemon #sp.def: The Base Special Defense of the Pokemon #speed: The Base Speed of the Pokemon

#generation: The numbered generation which the Pokemon was first introduced #legendary: Denotes if the Pokemon is legendary.

###

### 1. Import and clean data # Import data from csv

pokemon <- read.csv(file = "pokemon.csv", stringsAsFactors = FALSE) #this is done to import the data in R.

#FALSE imlies that at this stage all the variables will be stored as vectors. #However, later categorical variables should be stored as factors.

# str() shows the structure of data str(pokemon)

# examine the number of rows and cols nrow(pokemon)#shows total number of rows ncol(pokemon)

# Show the head and tail rows of a data frame head(pokemon)# head will show first 6 lines of data frame.

pokemon[l:6,]#on the left side, specify rows (in this case 1 through 6), #on the right,after comma, we specify the number of columns. head(pokemon, n=l)#n=l will only give the first row of dataset. tail(pokemon)#tail will give the last 6 rows

# summary() shows the mean and the five-number statistics indicating the spread

#of each column's values

summary(pokemon)#will show summary statistics for all the variables. However, #if categorical variables are stored as vectors, then it will show meaningless #statistics of categorical variab les.

# Remove unique identifiers (pokemon number and Name) from further analysis. pokemon <- pokemon[,c(-l,-2)]#before transforming categorical variables into #numerical you first need to remove rows that do not provide useful info. In

#this case pokemon name and number do not provide useful info.

# Change categorical variables to factors str(pokemon)

pokemon$Type.l <- factor(pokemon$Type.l)#we assign Type.l factor into original #variable (vector)Type.l. We transform it this way.

#then we transform the rest of the categorical variables. pokemon$Type.2 <- factor(pokemon$Type.2) pokemon$Generation <- factor(pokemon$Generation) pokemon$Legendary <- factor(pokemon$Legendary) str(pokemon)

summary(pokemon)

summary(pokemon$Type.l)#this will give the summary of frequencies of all #categories in the Type.l variable.

# set missing values in Type.2 as none

levels(pokemon$Type.2)#this will show all of the unique categories (levels) #in Type.2 pokemon

levels(pokemon$Type.2)[l]#this will show the first level, in this case its #missing.

levels(pokemon$Type.2)[l] <- "none"#it is good practice to assign none to the

#missing values. str(pokemon) summary(pokemon)

### 2. understanding a single variable: numeric variables # Show summary of one or more columns

summary(pokemon$Attack)#this summary function, applied on numerical variable

#"Attack", shows summary statistics of this variable. summary(pokemon[,c("Attack", "Defense")])#summary of 2 numberic variables. We #need to use combine functio since there are more than 1 variables.

Alternative

#would be writing c(4,5)

summary(pokemon$Sp..Atk + pokemon$Sp..Def)#Summary info of a NEW variable which

#is a sum of special attack and special defence.

which(pokemon$Sp..Atk + pokemon$Sp..Def == 340)#will show pokemon that has a sum

#of special attack and defence= to 340

# obtain the mean, median, and range of a numeric variable mean(pokemon$Attack)

median(pokemon$Attack) range(pokemon$Attack)

# use quantile to calculate the five-number summary for Attack quantile(pokemon$Attack)

# IQR

IQR(pokemon$Attack)#difference between 1st and 3d quartile

# boxplot of numeric variables

boxplot(pokemon$Attac,k main="Boxplot of Attack in the pokemon data set", ylab="Attack")

boxplot(pokemon$Defens,e main="Boxplot of Defense in the pokemon data set", ylab="Defense")

boxplot(pokemon[which(pokemon$Generation==l),4 ], main="Boxplot of Attack of the 1st generation pokemon", ylab="Attack") boxplot(pokemon[which(pokemon$Generation==l),5 ], main="Boxplot of Defense of the 1st generation pokemon", ylab="Defense")

# histograms of a numeric variable

hist(pokemon$Attack, main= "Histogram of Attack in the pokemon data set", xlab = "Attack")

hist(pokemon$Defense, main= 'H'istogram of Defense in the pokemon data set", xlab = "Defense")

hist(pokemon$HP, main= "Histogram of HP in the pokemon data set", xlab = "HP")

# variance and standard deviation of a numeric varaible var(pokemon$Attack)

sd(pokemon$Attack)

### 3. Exploring categorical variables

# Summary of categorical variable summary(pokemon$Type.l)#gives count nlevels(pokemon$Type.l)#will tell how many unique categories

# Plot categorica l variable

plot(pokemon$Type.l, main= "Plot of Type.l in the pokemon data set", xlab "Type.l")

table(pokemon$Type.l)#wil l show table with frequency sort(table(pokemon$Type.l))

# Run prop.table

Type table = table(pokemon$Type.l)#THis creates a table for Type 1, you have #to create it before you can find proportions

prop.table(Type table)#this will show proportions

### 4. Understand relationships of multiple variables # scatter plot: two numeric variables

plot(pokemon$Attack, pokemon$Defense)#scatter plot of 2 numberic variables

# Generate correlation coefficients of two numeric variables in a **2x2** matrix

# cor(X,Y) lies between -1 and 1. zero means no correlation. 1 or -1 indicates full correlation

# positive value means positive correlation and negative values mean negative relationships

cor(pokemon[,c("Attack", "Defense")]) cor(pokemon[,c(4,5)])

# Generate the correlation matrix of all numeric variables cor(pokemon[,3:8])

# Generate 20 scatter plots

pairs(pokemon[,3:B] ) #scatterplots for columns 3 to 8

## Examine relationships between numeric variables and factors

# boxplot groups values of a numeric variable based on the values of a factor boxplot(Attack~Type.l, data pokemon)

boxplot(Attack~Type.l, data= pokemon[which(pokemon$Legendary=='True'),]) boxplot(Attack~Type.l, data= pokemon[which(pokemon$Legendary=='False'),]) boxplot(HP~Legendary, data= pokemon)

boxplot(HP~Type.l, data= pokemon)

# Part 2: CarAuction Data

###

# This dataset contains information of cars purchased at the Auction. # VARIABLE DESCRIPTIONS:

#Auction: Auction provider at which the vehicle was purchased #Color: Vehicle Color

#IsBadBuy: Identifies if the kicked vehicle was an avoidable purchase #MMRCurrentAuctionAveragePrice: Acquisition price for this vehicle in average condition as of current day

#Size: The size category of the vehicle (Compact, SUV, etc.) #TopThreeAmericanName:Identifies if the manufacturer is one of the top three American manufacturers

#VehBCost: Acquisition cost paid for the vehicle at time of purchase #VehicleAge: The Years elapsed since the manufacturer's year #VehOdo: The vehicles odometer reading

#WarrantyCost: Warranty price (term=36month and millage=36K) #WheelType: The vehicle wheel type description (Alloy, Covers) ###

### 1. Import and clean data # Import data from csv

carAuction <- read.csv(file = "carAuction.csv", stringsAsFactors = FALSE) ### I spend 3 hrs on the first part. I wrote a lot of comments. But every now #and then this software DISCONNECTS. first 5-6 times my comments were

#still there. But the last time it happened they were ALL gone.

# str() shows the structure of data str(carAuction)

# summary() shows the mean and the five-number statistics indicating the spread of each column's values

summary(carAuction)

# Change all categorical variables to factors #first, check which variables are categorical carAuction$Auction = factor(carAuction$Auction) carAuction$Color = factor(carAuction$Color) carAuction$IsBadBuy = factor(carAuction$IsBadBuy) carAuction$Size = factor(carAuction$Size)

carAuction$TopThreeAmericanName = factor(carAuction$TopThreeAmericanName) carAuction$WheelType = factor(carAuction$WheelType)

str(carAuction)

summary(carAuction)

### 2. understanding a single variable: numerical variables

# Show summary of VehOdo summary(carAuction$VehOdo)

# obtain the mean, median, and range of WarrantyCost mean(carAuction$WarrantyCost) median(carAuction$WarrantyCost) range(carAuction$WarrantyCost)

# use quantile to calculate the five-number summary for WarrantyCost quantile(carAuction$WarrantyCost)

# display the IQR of WarrantyCost IQR(carAuction$WarrantyCost)

# boxplot of numeric variables: VehBCost and VehicleAge boxplot(carAuction$VehBCos,t main="Boxplot of vehicle cost", ylab="Cost") boxplot(carAuction$VehicleAg,e main="Boxplot of vehicle age", ylab="Age") boxplot(VehBCost~VehicleAge, data= carAuction)

# histograms of VehOdo

hist(carAuction$VehOdo, main= "Histogram of VehOdo", xlab "VehOdo") ### 3. Exploring categorical variables

# Show the number of cars in different WheelType nlevels(carAuction$WheelType)

# Show the proportion of cars in different WheelType Type\_table = table(carAuction$WheelType) prop.table(Type\_table)

### 4. Understand relationships of multiple variables

# scatter plot: VehBCost and MMRCurrentAuctionAveragePrice plot(carAuction$VehBCos,t carAuction$MMRCurrentAuctionAveragePrice)

# Generate correlation coefficients of VehBCost and MMRCurrentAuctionAveragePrice

cor(carAuction[,c("VehBCost", "MMRCurrentAuct ionAveragePrice")])

# Generate the correlation matrix of all numeric variables (you can specify the numeric variable names or their column numbers in a c() function)

cor(carAuction[,c('MMRCurrentAuct ionAveragePrice','VehBCost','VehicleAge ','VehOdo', 'WarrantyCost ')])

## Examine relationships between numeric variables and factors # boxplot VehBCost based on IsBadBuy boxplot(VehBCost~IsBadBuy, data= carAuction)

# Question: list one thing you learned from the carAuction data exploration. #I learned to save my work in Ras I go. All of my comments for the first

#section were gone after software disconencted me. I will need to doing again to print it out and study.

#aslo: I learned that VehBCOst and MMRAverageprice are positively #corelated.

#also: I learned to write function to get summary statistics.

# Spent about 2hours on first section, since i was pausing and writing comments,

#but they are all gone :(