A00275664_RMarkdown_3

A00275664

2023-11-30

RMarkdown Document

Question 3

Importing and Analysing the Data

```
CarsDataset <- read.csv("C:/Users/erink/OneDrive/Desktop/Interpretation of Data/CarsDataset.csv")

names(CarsDataset)
```

```
##  [1] "id"           "url"          "region"       "region_url"   "price"
##  [6] "year"         "manufacturer" "model"        "condition"    "cylinders"
## [11] "fuel"         "odometer"     "title_status" "transmission" "VIN"
## [16] "drive"        "size"         "type"         "paint_color"  "image_url"
## [21] "description"  "state"        "lat"          "long"         "posting_date"
```

```
str(CarsDataset) #id, price, year, cylinders, odometer, lat, long, posting_date - wrong formats
```

Handling Null Values

```
CarsDataset1a<-CarsDataset
#Replaces Blank Spaces with NA, if any
CarsDataset1a <- mutate(CarsDataset1a,
                        id= na_if(id, ""),
                        url= na_if(url, ""),
                        region= na_if(region, ""),
                        region_url= na_if(region_url, ""),
                        price= na_if(price, ""),
                        year= na_if(year, ""),
                        manufacturer= na_if(manufacturer, ""),
                        model= na_if(model, ""),
                        condition= na_if(condition, ""),
                        cylinders= na_if(cylinders, ""),
                        fuel= na_if(fuel, ""),
                        odometer= na_if(odometer, ""),
                        title_status= na_if(title_status, ""),
                        transmission= na_if(transmission, ""),
```

```
                        VIN= na_if(VIN, ""),
                        drive= na_if(drive, ""),
                        size= na_if(size, ""),
                        type= na_if(type, ""),
                        paint_color= na_if(paint_color, ""),
                        image_url= na_if(image_url, ""),
                        description= na_if(description, ""),
                        state= na_if(state, ""),
                        lat= na_if(lat, ""),
                        long= na_if(long, ""),
                        posting_date= na_if(posting_date, ""))
```

Column total reduced from 42547 to 42446

THE ID COLUMN

All the columns in the Cars Dataset have data jumbled within it that is not relevent to its column To remove the data, I will convert it to a character then numeric form which will turn it into an NA value

```
#Convert to numeric
Cars_Dataset1b<-CarsDataset1a
Cars_Dataset1b$id <- as.numeric(as.character(Cars_Dataset1b$id))
```

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

```
# Remove NA values
Cars_Dataset1b<-filter(Cars_Dataset1b, !is.na(id))

#Left with 42400 rows of data
```

THE URL COLUMN

```
# Using stringr, remove all values that do not begin with "https://
Cars_Dataset1b <- Cars_Dataset1b[str_detect(Cars_Dataset1b$url, "https://"), ]

#Changing all values to lower case
Cars_Dataset1b$url <- str_to_lower(Cars_Dataset1b$url)

# Remove all NA values
Cars_Dataset1b<-filter(Cars_Dataset1b, !is.na(url))

#Left with 42399 rows of data
```

THE REGION COLUMN

```
# Changing all values to lower case
Cars_Dataset1b$region <- str_to_lower(Cars_Dataset1b$region)

# Remove all NA values
Cars_Dataset1b<-filter(Cars_Dataset1b, !is.na(region))
```

THE REGION_URL COLUMN

```
# Changing all values to lower case
Cars_Dataset1b$region_url <- str_to_lower(Cars_Dataset1b$region_url)

# Using stringr, remove all values that do not begin with "https://
Cars_Dataset1b <- Cars_Dataset1b[str_detect(Cars_Dataset1b$region_url, "https://"), ]

# Remove all NA values
Cars_Dataset1b<-filter(Cars_Dataset1b, !is.na(region_url))
```

THE PRICE COLUMN

```
# Change to numeric
Cars_Dataset1b$price <- as.numeric(as.character(Cars_Dataset1b$price))

#Remove prices that are 0
Cars_Dataset1b <- Cars_Dataset1b %>%
  filter(price != 0)

# Remove NA values
Cars_Dataset1b<-filter(Cars_Dataset1b, !is.na(price))

#Left with 39727 rows of data
```

THE YEAR COLUMN

```
# Change to numeric for better analysis
Cars_Dataset1b$year <- as.numeric(as.character(Cars_Dataset1b$year))

#Filter any year that is a 0, if any
Cars_Dataset1b <- Cars_Dataset1b %>%
  filter(year != 0)

# Remove NA values
Cars_Dataset1b<-filter(Cars_Dataset1b, !is.na(year))
```

THE MANUFACTURER COLUMN

```
# Changing all values to lower case
Cars_Dataset1b$manufacturer <- str_to_lower(Cars_Dataset1b$manufacturer)

# Replacing similiar condition variables
Cars_Dataset1b <- Cars_Dataset1b %>%
  mutate(manufacturer = case_when(manufacturer == "rover" ~ "jaguar",
                                  manufacturer == "land rover" ~ "jaguar",
                                  manufacturer == "jeep" ~ "chrysler",
                                  manufacturer == "acura" ~ "honda",
                                  manufacturer == "dodge" ~ "chrysler",
                                  manufacturer == "ram" ~ "chrysler",
                                  manufacturer == "lexus" ~ "toyota",
                                  manufacturer == "infinit" ~ "nissan",
                                  manufacturer == "mini" ~ "bmw",
                                  manufacturer == "datsun" ~ "nissan",
                                  manufacturer == "chevrolet" ~ "gmc",
                                  manufacturer == "pontiac" ~ "gmc",
                                 TRUE ~ as.character(manufacturer)))

# Remove all NA values
Cars_Dataset1b<-filter(Cars_Dataset1b, !is.na(manufacturer))
```

THE MODEL COLUMN

```
# Changing all values to lower case
Cars_Dataset1b$model <- str_to_lower(Cars_Dataset1b$model)

# Remove all NA values
Cars_Dataset1b<-filter(Cars_Dataset1b, !is.na(model))
```

THE CONDITION COLUMN

```
# Changing all values to lower case
Cars_Dataset1b$condition <- str_to_lower(Cars_Dataset1b$condition)

# Replacing similiar condition variables
Cars_Dataset1b <- Cars_Dataset1b %>%
  mutate(condition = case_when(condition == "like new" ~ "good",
                               condition == "excellent" ~ "excellent",
                               condition == "good" ~ "good",
                               condition == "fair" ~ "fair",
                               condition == "salvage" ~ "salvage",
                               condition == "new" ~ "new",
                               TRUE ~ as.character(condition)))

# Remove all NA values
Cars_Dataset1b<-filter(Cars_Dataset1b, !is.na(condition))
```

THE CYLINDERS COLUMN

```
# Changing all values to lower case
Cars_Dataset1b$cylinders <- str_to_lower(Cars_Dataset1b$cylinders)

# Replacing cylinder variables with numeric variables
Cars_Dataset1b <- Cars_Dataset1b %>%
  mutate(cylinders = case_when(cylinders == "6 cylinders" ~ "6",
                               cylinders == "4 cylinders" ~ "4",
                               cylinders == "8 cylinders" ~ "8",
                               cylinders == "10 cylinders" ~ "10",
                               cylinders == "3 cylinders" ~ "3",
                               cylinders == "six cylinders" ~ "6",
                               cylinders == "5 cylinders" ~ "5",
                               cylinders == "12 cylinders" ~ "12",
                               cylinders == "8 cyls" ~ "8",
                               cylinders == "6 cyls" ~ "6",
                               cylinders == "4 cyls" ~ "4",
                               TRUE ~ as.character(cylinders)))


# Converting the column to a Numeric data type
Cars_Dataset1b$cylinders<- as.numeric(as.character(Cars_Dataset1b$cylinders))
```

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

```
# Remove all NA values created
Cars_Dataset1b<-filter(Cars_Dataset1b, !is.na(cylinders))

#Left with 39631 rows of data
```

THE FUEL COLUMN

```
# Changing all values to lower case
Cars_Dataset1b$fuel <- str_to_lower(Cars_Dataset1b$fuel)

# Remove all NA values created
Cars_Dataset1b<-filter(Cars_Dataset1b, !is.na(fuel))
```

THE ODOMETER COLUMN

```
# Converting the column to a Numeric data type
Cars_Dataset1b$odometer<- as.numeric(as.character(Cars_Dataset1b$odometer))

# Remove all NA values created
Cars_Dataset1b<-filter(Cars_Dataset1b, !is.na(odometer))
```

THE TITLE_STATUS COLUMN

```
# Changing all values to lower case
Cars_Dataset1b$title_status <- str_to_lower(Cars_Dataset1b$title_status)

# Remove all NA values created
Cars_Dataset1b<-filter(Cars_Dataset1b, !is.na(title_status))

###------INVESTIGATE UNIQUE VARIABLES-----######
```

THE TRANSMISSION COLUMN

```
# Changing all values to lower case
Cars_Dataset1b$transmission <- str_to_lower(Cars_Dataset1b$transmission)

# Remove all NA values created
Cars_Dataset1b<-filter(Cars_Dataset1b, !is.na(transmission))
```

THE VIN COLUMN #39613

There is too much data to search the VIN column for abstract data. I will filter out any cells that contain 0 values and null values

```
Cars_Dataset1b <- Cars_Dataset1b %>%
  filter(VIN != 0)
# Remove all NA values created
Cars_Dataset1b<-filter(Cars_Dataset1b, !is.na(VIN))
```

THE DRIVE COLUMN

```
# Changing all values to lower case
Cars_Dataset1b$drive <- str_to_lower(Cars_Dataset1b$drive)

# Replacing similiar variables with consistent variables
Cars_Dataset1b <- Cars_Dataset1b %>%
  mutate(drive = case_when(drive == "4wd" ~ "four wheel drive",
                           drive == "rwd" ~ "rear wheel drive",
                           drive == "fwd" ~ "front wheel drive",
                           drive == "front wheel drive" ~ "front wheel drive",
                           drive == "rear wheel drive" ~ "rear wheel drive",
                           drive == "four wheel drive" ~ "four wheel drive",
                           TRUE ~ as.character(drive)))

# Remove all NA values created
Cars_Dataset1b<-filter(Cars_Dataset1b, !is.na(drive))
```

THE SIZE COLUMN

```
# Changing all values to lower case
Cars_Dataset1b$size <- str_to_lower(Cars_Dataset1b$size)

# Remove all NA values created
Cars_Dataset1b<-filter(Cars_Dataset1b, !is.na(size))
```

THE TYPE COLUMN

```
# Changing all values to lower case
Cars_Dataset1b$type <- str_to_lower(Cars_Dataset1b$type)

# Remove all NA values
Cars_Dataset1b<-filter(Cars_Dataset1b, !is.na(type))
```

THE PAINT_COLOR COLUMN

```
# Changing all values to lower case
Cars_Dataset1b$paint_color <- str_to_lower(Cars_Dataset1b$paint_color)

# Remove all NA values
Cars_Dataset1b<-filter(Cars_Dataset1b, !is.na(paint_color))
```

THE IMAGE_URL COLUMN

```
# Using stringr, remove all values that do not begin with "https://
Cars_Dataset1b <- Cars_Dataset1b[str_detect(Cars_Dataset1b$image_url, "https://"), ]

# Remove all NA values
Cars_Dataset1b<-filter(Cars_Dataset1b, !is.na(image_url))
```

THE DESCRIPTION COLUMN

```
# Remove all NA values
Cars_Dataset1b<-filter(Cars_Dataset1b, !is.na(description))
```

THE STATE COLUMN

```
# Changing all values to lower case
Cars_Dataset1b$state <- str_to_lower(Cars_Dataset1b$state)

# Remove all NA values
Cars_Dataset1b<-filter(Cars_Dataset1b, !is.na(state))
```

THE LAT COLUMN

```
# Converting the column to a Numeric data type
Cars_Dataset1b$lat<- as.numeric(as.character(Cars_Dataset1b$lat))

# Remove all NA values created
Cars_Dataset1b<-filter(Cars_Dataset1b, !is.na(lat))
```

THE LONG COLUMN

```
# Converting the column to a Numeric data type
Cars_Dataset1b$long<- as.numeric(as.character(Cars_Dataset1b$long))

# Remove all NA values
Cars_Dataset1b<-filter(Cars_Dataset1b, !is.na(long))
```

THE POSTING_DATE

```
# Seperate the column into three for a more detailed analysis
Cars_Dataset1b <- separate(Cars_Dataset1b, posting_date, into = c("date", "time_zone"), sep = "T")
Cars_Dataset1b <- separate(Cars_Dataset1b, time_zone, into = c("time", "timezone"), sep = "-")
```

##THE DATE COLUMN

```
# Convert the data format to date using lubridate

Cars_Dataset1b$date<- ymd(Cars_Dataset1b$date)

# Remove all NA values
Cars_Dataset1b<-filter(Cars_Dataset1b, !is.na(date))
```

THE TIME COLUMN

```
# Convert the character format of time to time format(hms)

Cars_Dataset1b$time <- hms(Cars_Dataset1b$time)

# Remove all NA values
Cars_Dataset1b<-filter(Cars_Dataset1b, !is.na(time))
```

THE TIMEZONE COLUMN

```
# Convert the data format to numeric format

Cars_Dataset1b$timezone<- as.numeric(as.character(Cars_Dataset1b$timezone))

# Remove all NA values
Cars_Dataset1b<-filter(Cars_Dataset1b, !is.na(timezone))
```

Remove Duplicate Data

```
# Removing all duplicates

Cars_Dataset1b<- unique(Cars_Dataset1b) # The dataset reduced from 39613 to 39598 rows
```