Data in Motion, LLC: WeeK#19

url <- 'https://raw.githubusercontent.com/kedeisha1/Challenges/main/coaster\_db.csv'  
data<- read.csv(url)

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ forcats 1.0.0 ✔ readr 2.1.4  
## ✔ ggplot2 3.4.2 ✔ 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 ]8;;http://conflicted.r-lib.org/conflicted package]8;; to force all conflicts to become errors

library(lubridate)  
library(ggplot2)  
library(tidyr)  
  
  
#Question 1: How many columns and rows are in the dataset?  
nrow(data)# number of rows

## [1] 1087

ncol(data)# number of columns

## [1] 56

dim(data)

## [1] 1087 56

#Question 2: Is there any missing data?  
missing\_data <- data %>%  
 replace(is.na(.), "") %>% # replace NA values with empty string  
 summarise\_all(~sum(. == "")) # count missing values for each variable  
  
missing\_data <- missing\_data %>%  
 gather(variable, count\_missing) %>%  
 arrange(desc(count\_missing))  
  
missing\_data

## variable count\_missing  
## 1 Fastrack.available 1068  
## 2 Restraint.Style 1065  
## 3 Restraints 1063  
## 4 Opened 1060  
## 5 Name 1052  
## 6 Theme 1043  
## 7 Flash.Pass.available 1041  
## 8 Flash.Pass.Available 1037  
## 9 Acceleration 1027  
## 10 Fast.Lane.available 1018  
## 11 Single.rider.line.available 1006  
## 12 Website 1000  
## 13 Replaced.by 999  
## 14 Soft.opening.date 991  
## 15 Soft.opening.date.1 991  
## 16 Must.transfer.from.wheelchair 981  
## 17 height\_ft 916  
## 18 Replaced 914  
## 19 Closing.date 851  
## 20 Track.layout 752  
## 21 Max.vertical.angle 730  
## 22 G.force 725  
## 23 Gforce\_clean 725  
## 24 Cost 705  
## 25 Park.section 600  
## 26 Drop 593  
## 27 Capacity 512  
## 28 Designer 509  
## 29 Trains 369  
## 30 Model 343  
## 31 Duration 322  
## 32 Lift.launch.system 292  
## 33 latitude 275  
## 34 longitude 275  
## 35 Height.restriction 256  
## 36 Opening.date 250  
## 37 opening\_date\_clean 250  
## 38 Status 213  
## 39 Inversions 155  
## 40 speed2 152  
## 41 Speed 150  
## 42 speed1 150  
## 43 speed1\_value 150  
## 44 speed1\_unit 150  
## 45 speed\_mph 150  
## 46 Length 134  
## 47 Height 122  
## 48 height\_value 122  
## 49 height\_unit 122  
## 50 Manufacturer 59  
## 51 coaster\_name 0  
## 52 Location 0  
## 53 Type 0  
## 54 year\_introduced 0  
## 55 Type\_Main 0  
## 56 Inversions\_clean 0

sum(missing\_data$count\_missing)

## [1] 29435

In total, there are 29435 missing values and they sorted in descending order

#Question 3: Display the summary statistics of the numeric columns using the describe method.  
library(psych)# describe function is from psych package

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

##   
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

describe(data[, sapply(data, is.numeric)])# sapply used for numeric variabels

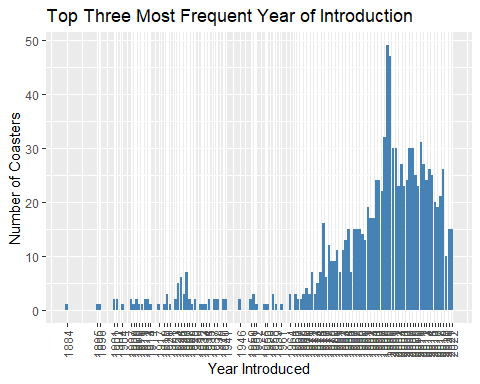
## vars n mean sd median trimmed mad min max  
## Inversions 1 932 1.55 2.11 0.00 1.18 0.00 0.00 14.00  
## year\_introduced 2 1087 1994.99 23.48 2000.00 1999.14 16.31 1884.00 2022.00  
## latitude 3 812 38.37 15.52 40.29 40.44 7.69 -48.26 63.23  
## longitude 4 812 -41.60 72.29 -76.65 -54.06 26.44 -123.04 153.43  
## speed1\_value 5 937 53.85 23.39 50.00 51.60 17.79 5.00 240.00  
## speed\_mph 6 937 48.62 16.68 49.70 48.16 15.27 5.00 149.10  
## height\_value 7 965 89.58 136.25 79.00 79.90 51.89 4.00 3937.00  
## height\_ft 8 171 102.00 67.33 91.20 93.42 59.30 13.10 377.30  
## Inversions\_clean 9 1087 1.33 2.03 0.00 0.92 0.00 0.00 14.00  
## Gforce\_clean 10 362 3.82 0.99 4.00 3.85 0.74 0.80 12.00  
## range skew kurtosis se  
## Inversions 14.00 1.34 1.56 0.07  
## year\_introduced 138.00 -2.01 4.42 0.71  
## latitude 111.49 -2.80 10.23 0.54  
## longitude 276.46 1.33 0.87 2.54  
## speed1\_value 235.00 1.76 7.39 0.76  
## speed\_mph 144.10 0.58 2.21 0.54  
## height\_value 3933.00 23.42 656.16 4.39  
## height\_ft 364.20 1.40 2.60 5.15  
## Inversions\_clean 14.00 1.56 2.24 0.06  
## Gforce\_clean 11.20 1.13 12.17 0.05

# Question 4:Rename the following columns:  
# a. coaster\_name ➡️ Coaster\_Name  
# b. year\_introduced ➡️ Year\_Introduced  
# c. opening\_date\_clean ➡️ Opening\_Date  
# d. speed\_mph ➡️ Speed\_mph  
# e. height\_ft ➡️ Height\_ft  
# f. Inversions\_clean ➡️ Inversions  
# g. Gforce\_clean ➡️ Gforce  
  
data <- data %>%   
 rename(Coaster\_Name = coaster\_name,   
 Year\_Introduced = year\_introduced,   
 Opening\_Date = opening\_date\_clean,   
 Speed\_mph = speed\_mph,   
 Height\_ft = height\_ft,   
 Inversion\_status = Inversions\_clean, # "Inversion" already existed in the dataframe, thus, it renamed as "Inversion\_status"  
 Gforce = Gforce\_clean)

# # Question 5: Are there any duplicated rows?  
data %>%   
 filter(duplicated(.))

## [1] Coaster\_Name Length   
## [3] Speed Location   
## [5] Status Opening.date   
## [7] Type Manufacturer   
## [9] Height.restriction Model   
## [11] Height Inversions   
## [13] Lift.launch.system Cost   
## [15] Trains Park.section   
## [17] Duration Capacity   
## [19] G.force Designer   
## [21] Max.vertical.angle Drop   
## [23] Soft.opening.date Fast.Lane.available   
## [25] Replaced Track.layout   
## [27] Fastrack.available Soft.opening.date.1   
## [29] Closing.date Opened   
## [31] Replaced.by Website   
## [33] Flash.Pass.Available Must.transfer.from.wheelchair  
## [35] Theme Single.rider.line.available   
## [37] Restraint.Style Flash.Pass.available   
## [39] Acceleration Restraints   
## [41] Name Year\_Introduced   
## [43] latitude longitude   
## [45] Type\_Main Opening\_Date   
## [47] speed1 speed2   
## [49] speed1\_value speed1\_unit   
## [51] Speed\_mph height\_value   
## [53] height\_unit Height\_ft   
## [55] Inversion\_status Gforce   
## <0 rows> (or 0-length row.names)

# # Question 6: What are the top 3 years with the most roller coasters introduced?  
  
  
top\_three <- data %>%   
 group\_by(Year\_Introduced) %>%   
 count()   
  
ggplot(top\_three, aes(x = Year\_Introduced, y = n)) +  
 geom\_bar(stat = "identity", fill = "steelblue") +  
 labs(x = "Year Introduced", y = "Number of Coasters") +  
 ggtitle("Top Three Most Frequent Year of Introduction")+  
 scale\_x\_continuous( breaks = unique(top\_three$Year\_Introduced))+ # to display year on x-axis   
 theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1)) # to make "year" vertical on plot



*top three years* are 1998, 1999, 2000

# What is the average speed? Also display a plot to show it’s distribution.