# **HW 05 - Legos**

library(tidyverse)

library(dsbox)

1. What are the three most common first names of purchasers?

lego\_sales %>%

count(first\_name, sort = TRUE)

## # A tibble: 211 x 2

## first\_name n

## <chr> <int>

## 1 Jackson 13

## 2 Jacob 11

## 3 Joseph 11

## 4 Michael 10

## 5 Audrey 8

## 6 Connor 8

## 7 Kaitlyn 8

## 8 Lucas 8

## 9 Amanda 7

## 10 Joshua 7

## # … with 201 more rows

✅ In this sample, the three common first names of purchasers are *Jackson, Jacob, Joseph*

1. What are the three most common themes of Lego sets purchased?

lego\_sales %>%

count(theme, sort = TRUE)

## # A tibble: 25 x 2

## theme n

## <chr> <int>

## 1 Star Wars 75

## 2 Nexo Knights 64

## 3 Gear 55

## 4 Mixels 55

## 5 City 45

## 6 Friends 42

## 7 Ninjago 38

## 8 Duplo 35

## 9 Bionicle 34

## 10 Creator 25

## # … with 15 more rows

✅ In this sample, the three most common themes of Lego sets purchased are *Star Wars, Nexo Knights, Gear, and Mixels*

1. Among the most common theme of Lego sets purchased, what is the most common subtheme?

lego\_sales %>%

filter(theme == "Star Wars") %>%

count(subtheme, sort = TRUE)

## # A tibble: 11 x 2

## subtheme n

## <chr> <int>

## 1 The Force Awakens 15

## 2 Buildable Figures 11

## 3 Episode V 10

## 4 MicroFighters 10

## 5 Battlefront 7

## 6 Original Content 7

## 7 Episode III 6

## 8 Rebels 3

## 9 Seasonal 3

## 10 Episode IV 2

## 11 Ultimate Collector Series 1

✅ In this sample, the most common theme of Lego sets purchased is *Star Wars*, and the most common subtheme is *The Force Awakens*

1. Create a new variable called age\_group and group the ages into the following categories: “18 and under”, “19 - 25”, “26 - 35”, “36 - 50”, “51 and over”.

lego\_sales <- lego\_sales %>%

mutate(age\_group = case\_when(

age <= 18 ~ "18 and under",

age >= 19 & age <= 25 ~ "19 - 25",

age >= 26 & age <= 35 ~ "26 - 35",

age >= 36 & age <= 50 ~ "36 - 50",

age >= 51 ~ "51 and over"

))

1. Which age group has purchased the highest number of Lego sets.

lego\_sales %>%

count(age\_group, sort = TRUE)

## # A tibble: 5 x 2

## age\_group n

## <chr> <int>

## 1 36 - 50 216

## 2 26 - 35 183

## 3 19 - 25 129

## 4 51 and over 62

## 5 18 and under 30

✅ In this sample group of *36-50* yo purchased the highest number of Lego sets.

1. Which age group has spent the most money on Legos?

lego\_sales %>%

mutate(

amount\_spent = us\_price \* quantity

) %>%

group\_by(age\_group) %>%

summarise(

total\_spent = sum(amount\_spent)

) %>%

arrange(desc(total\_spent))

## # A tibble: 5 x 2

## age\_group total\_spent

## <chr> <dbl>

## 1 36 - 50 9533.

## 2 26 - 35 7576.

## 3 19 - 25 4939.

## 4 51 and over 2475.

## 5 18 and under 949.

✅ In this sample group of *36-50* yo has spent the most money on Legos.

1. Which Lego theme has made the most money for Lego?

lego\_sales %>%

mutate(

amount\_spent = us\_price \* quantity

) %>%

group\_by(theme) %>%

summarise(

total\_spent = sum(amount\_spent)

) %>%

arrange(desc(total\_spent))

## # A tibble: 25 x 2

## theme total\_spent

## <chr> <dbl>

## 1 Star Wars 4448.

## 2 Ninjago 2279.

## 3 City 2211.

## 4 Nexo Knights 2209.

## 5 Minecraft 1550.

## 6 Gear 1533.

## 7 Friends 1279.

## 8 Duplo 1220.

## 9 Elves 1120.

## 10 Ghostbusters 880.

## # … with 15 more rows

✅ In this sample *Star Wars* has made the most money for Lego

1. Which area code has spent the most money on Legos? In the US the area code is the first 3 digits of a phone number.

lego\_sales %>%

mutate(

amount\_spent = us\_price \* quantity

) %>%

group\_by(substr(phone\_number, 1, 3)) %>%

summarise(

total\_spent = sum(amount\_spent)

) %>%

arrange(desc(total\_spent))

## # A tibble: 157 x 2

## `substr(phone\_number, 1, 3)` total\_spent

## <chr> <dbl>

## 1 <NA> 3993.

## 2 956 720.

## 3 973 685.

## 4 567 550.

## 5 281 465.

## 6 316 438.

## 7 339 426.

## 8 209 350.

## 9 423 340.

## 10 778 335.

## # … with 147 more rows

✅ In this sample area code of *956 (state of Texas)* has spent the most money on Legos

1. Come up with a question you want to answer using these data, and write it down. Then, create a data visualization that answers the question, and explain how your visualization answers the question.

lego\_sales %>%

count(theme, age\_group) %>%

group\_by(theme) %>%

mutate(proc = n / sum(n)) %>%

ggplot(aes(y =theme, x = proc, fill = age\_group)) +

geom\_col(position = 'dodge') +

theme\_minimal() +

labs(

title = "Which age group is most commonly to buy a specific theme of Lego?",

subtitle = "Lego sales in 2018 for a sample of customers who bought Legos in the US.",

x = NULL, y = NULL, fill = "Age group"

) +

scale\_x\_continuous(label = label\_percent())

