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*

2. 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*

3. 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*

4. 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"
  ))
```

5. 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.

6. 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.

7. 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

8. 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.

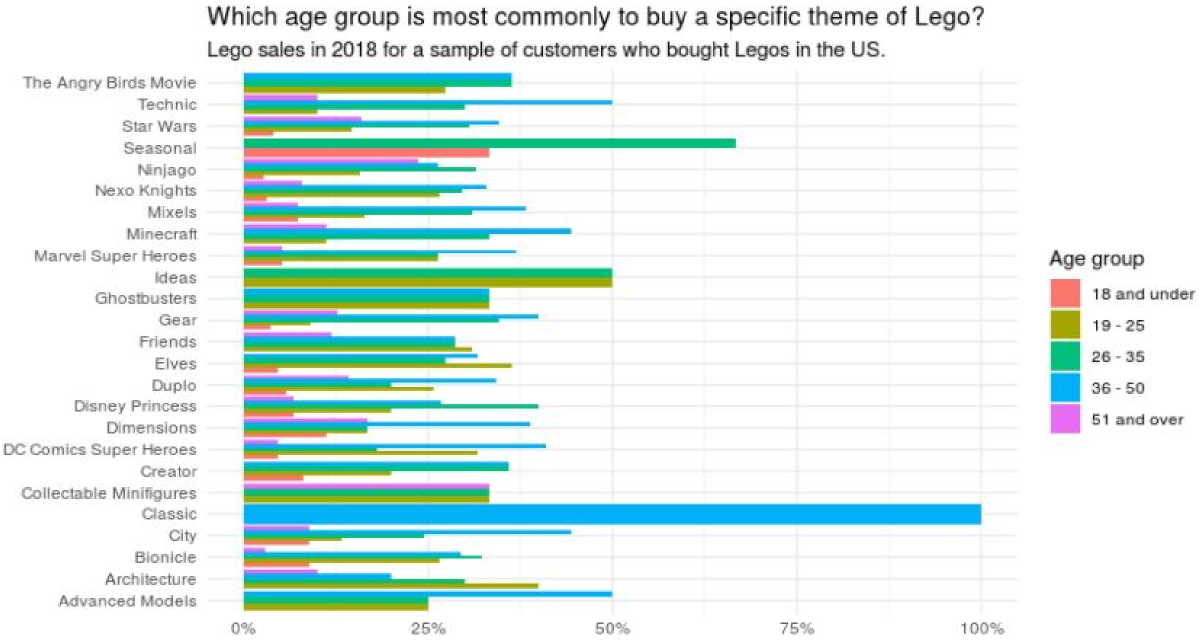
```
lego_sales %>%
  mutate(
    amount_spent = us_price * quantity
  ) %>%
  group_by(substr(phone_number, 1, 3)) %>%
  summarise(
    total_spent = sum(amount_spent)
  ) %>%
  arrange(desc(total_spent))
## # A tibble: 157 x 2
##    `substr(phone_number, 1, 3)` total_spent
##    <chr>                              <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

9. 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.

Q: Which age group is most commonly to buy a specific theme of Lego?

```
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())
```



If we take a look at the graph we may find which age group is more likely (commonly) to buy a specific theme of Lego. For example, in the "Ghostbusters" theme - where we can see that according to this data set, this theme is more likely to be bought by three different age groups (19-25, 26-35, 36-50), as all of these groups were the only buyers of this specific theme the chance that the buyer will be from this three groups are ~33.3%.
Another good example is the "Classic" theme in this data set. We may find that the only group who bought it, was 36-50 years old, so according to this data set, we might say - Age group of 36-50 years old are the only buyers of a "Classic" Lego theme.