Lego Homework - Rumi Allbert

Code

Hide

```
library(tidyverse)
library(dsbox)
```

Exercises

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

Hide

```
name_count <- lego_sales %>%
     count(first_name)

sorted_name_count <- name_count %>%
     arrange(desc(n))

sorted_name_count[1:3,]
```

first_name <chr>	n <int>
Jackson	13
Jacob	11
Joseph	11

3 rows

In this sample, the most common first names among the purchasers are: Jackson (13), Jacob (11) and Jospeph (11).

Exercise 2: What are the three most common themes of Lego sets purchased?

Hide

```
theme_count <- lego_sales %>%
     count(theme)

sorted_theme_count <- theme_count %>%
     arrange(desc(n))

sorted_theme_count[1:3,]
```

theme <chr>	n <int>
Star Wars	75
Nexo Knights	64
Gear	55

3 rows

In this sample, the most common themes of lego among the purchasers are: Star Wars (75), Nexo Knights (64) and Gear (55).

Exercise 3: Among the most common theme of Lego sets purchased, what is the most common subtheme?

Hide

```
most_commom_subtheme <- lego_sales %>%
  filter(theme == "Star Wars") %>%
  count(subtheme, sort = TRUE)
most_commom_subtheme[1,]
```

subtheme <chr>	n <int>
The Force Awakens	15

1 row

In this sample, the most common subtheme of the most common theme (Star Wars) is The Force Awakens (15).

Exercise 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”

Hide

```
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"))
```

For this exercise I create a new column in the lego sales dataframe which groups the different groups of ages, utilizing case_when().

Exercise 5: Which age group has purchased the highest number of Lego sets?

Hide

```
highest_n_legoset <- lego_sales %>%
    count(age_group, sort = TRUE)
highest_n_legoset[1,]
```

age_group <chr>	n <int>
36 - 50	216

1 row

For this sample, it is the age group of 36 - 50 who have purchased the higest number of lego sets.

Exercise 6: Which age group has spent the most money on Legos?

Hide

```
highest_spender_agegroup <- lego_sales %>%
    count(us_price, sort = TRUE)
highest_spender_agegroup[1,]
```

us_price <dbl>	n <int>
9.99	108

1 row

For this sample, it is the age group of 36 - 50 who have purchased the higest number of lego sets.

Exercise 7: Which Lego theme has made the most money for Lego?

Hide

```
lego_sales <- lego_sales %>%
    mutate(theme_income = us_price * quantity)

highest_income_theme <- lego_sales %>%
    group_by(theme, theme_income) %>%
    count(theme, sort = TRUE) %>%
    arrange(desc(theme_income))

highest_income_theme[1,]
```

theme <chr>	theme_income <dbl>	n <int>
Ghostbusters	699.98	1

1 row

Hide

```
NA
```

For this sample, it is the theme of Ghostbusters that generated the most money for Lego, acrruing an income of $699.98

Exercise 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.

Hide

```
lego_sales <- lego_sales %>%
    mutate(theme_income = us_price * quantity)

lego_sales <- lego_sales %>%
    mutate(area_code = str_sub((phone_number), 1, 3))

highest_income_areacode <- lego_sales %>%
    group_by(area_code, theme_income) %>%
    count(area_code, sort = TRUE) %>%
    arrange(desc(theme_income))

highest_income_areacode[1,]
```

area_code <chr>	theme_income <dbl>	n <int>
956	699.98	1

1 row

Hide

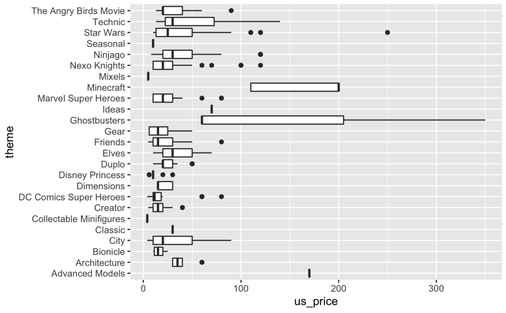
```
NA
```

For this sample, it is the area code 956 that generated the most money for Lego, acrruing an income of $699.98

Exercise 9: Why is Ghostbuster the lego theme that generated the most income? Is there a reason for this?

Hide

```
ggplot(lego_sales, aes(us_price, theme)) + geom_boxplot()
```



Creating a simle boxplot, it becomes very obvious tha there is an outlier for the Ghostbuster theme, which would skew the data. While all the other lego themes have a similar price range, the Ghostbuster theme has a very distant outlier. There are many reasons why this could be the case, perhaps the Ghostbuster theme is a limited model that goes for a high price.