Lego Sets Market Analysis and Price Prediction

Agne Adukaite Aklifazla
Mentor: Ricardo D. Alanis-Tamez

Some LEGO Facts

- Established by Danish carpenter Ole Kirk Christiansen in 1932 as a wooden interlocking toys.
- The Danish phrase leg godt [lɑj̝ˀ ˈkʌt] means "play well".
- In 1947, Lego expanded to begin producing plastic toys.
- The Lego Group's motto, "only the best is good enough"was created in 1936.
- The Lego Group's Duplo product line was introduced in 1969.
- In 1978, Lego produced the first minifigures.
- In May 2013, the largest model ever created was displayed in

New York City and was made of over 5 million bricks;

a 1:1 scale model of an X-wing fighter.

Data Wrangling

1. Missing values in review category:
- number of reviews,
- star rating,
- value star rating,
- review difficulty.

2. Data containing 1 piece.
3. Data not evenly distributed between countries.

Data Wrangling

Unique values information:

```
replace_values = {'Very Easy': 1, 'Easy': 2, 'Average' :3, 'Challenging': 4, 'Very Challenging':5}

legodata = legodata.replace({'review_difficulty': replace_values})
```

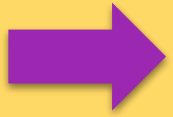
```
legodata["ages"].unique()
```

```
array(['6-12', '12+', '7-12', '10+', '8-12', '5-12', '4-99', '4+', '9-12',
       '16+', '14+', '9-14', '7-14', '8-14', '4-7', '6+', '2-5', '1½-5',
       '1½-3', '9+', '5-8', '8+', '6-14', '5+', '10-16', '10-14', '11-16',
       '12-16', '9-16', '7+'], dtype=object)
```

Data Wrangling

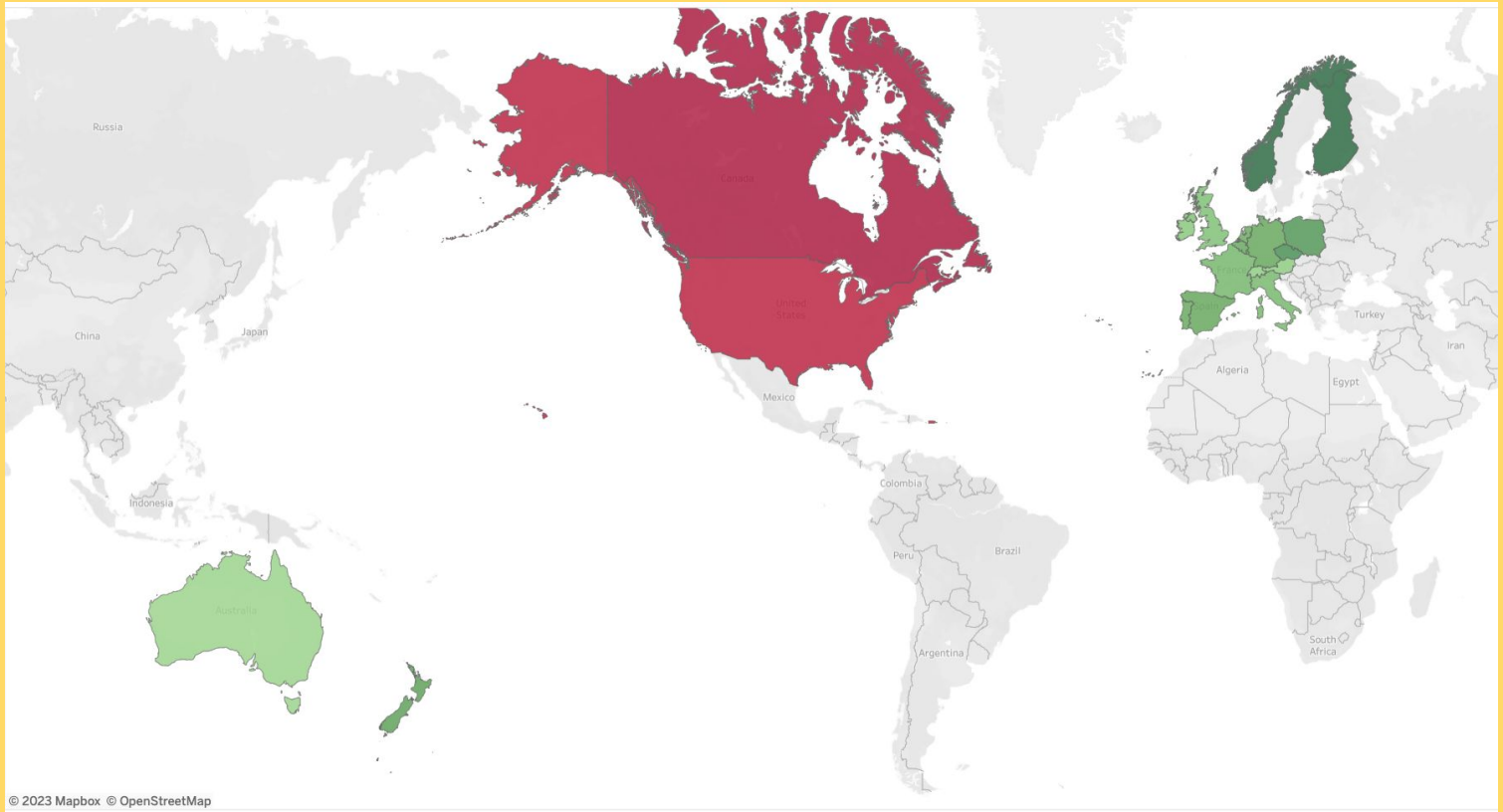
From 12261 entries to 9910 clean entries

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12261 entries, 0 to 12260
Data columns (total 14 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   ages               12261 non-null  object
 1   list_price         12261 non-null  float64
 2   num_reviews        10641 non-null  float64
 3   piece_count        12261 non-null  float64
 4   play_star_rating   10486 non-null  float64
 5   prod_desc          11884 non-null  object
 6   prod_id            12261 non-null  float64
 7   prod_long_desc     12261 non-null  object
 8   review_difficulty  10206 non-null  object
 9   set_name           12261 non-null  object
 10  star_rating        10641 non-null  float64
 11  theme_name         12258 non-null  object
 12  val_star_rating    10466 non-null  float64
 13  country            12261 non-null  object
dtypes: float64(7), object(7)
```

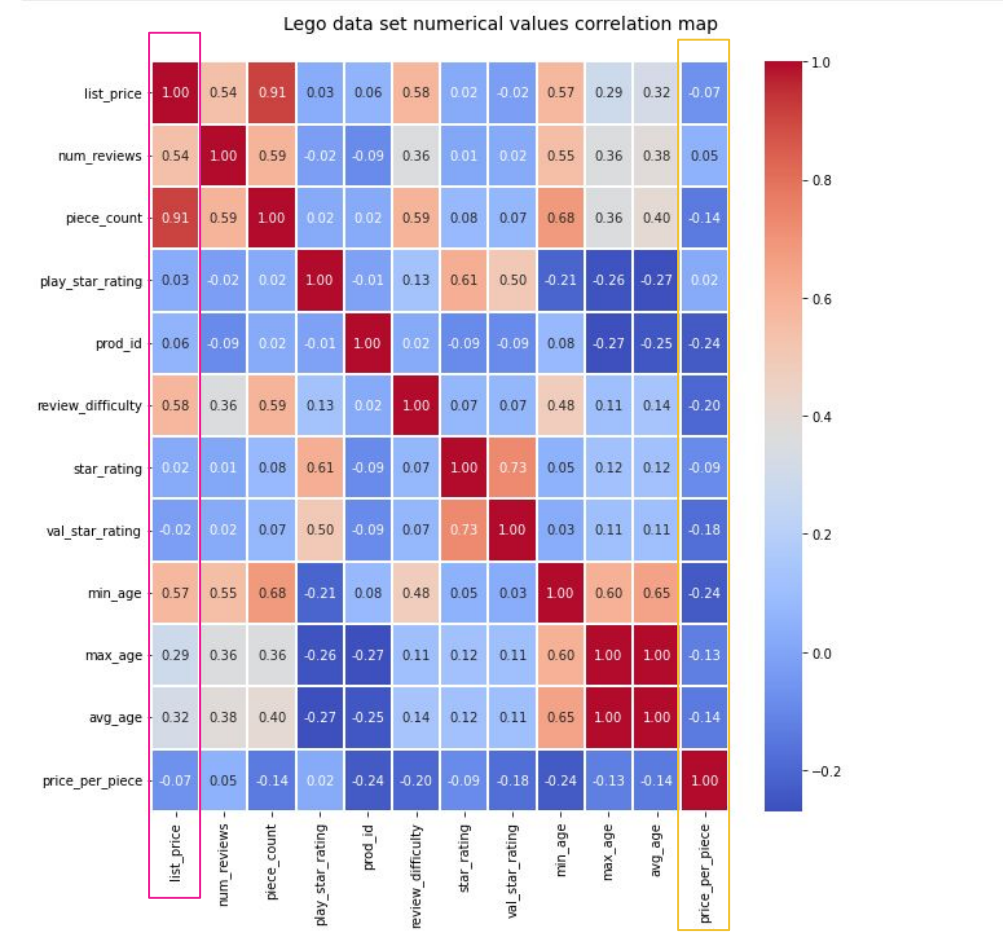


```
Int64Index: 9910 entries, 0 to 12260
Data columns (total 16 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   ages               9910 non-null   object
 1   list_price         9910 non-null   float64
 2   num_reviews        9910 non-null   float64
 3   piece_count        9910 non-null   float64
 4   play_star_rating   9910 non-null   float64
 5   prod_desc          9910 non-null   object
 6   prod_id            9910 non-null   float64
 7   prod_long_desc     9910 non-null   object
 8   review_difficulty  9910 non-null   int64
 9   set_name           9910 non-null   object
 10  star_rating        9910 non-null   float64
 11  theme_name         9910 non-null   object
 12  val_star_rating    9910 non-null   float64
 13  country            9910 non-null   object
 14  min_age            9910 non-null   float64
 15  max_age            9910 non-null   float64
dtypes: float64(9), int64(1), object(6)
```

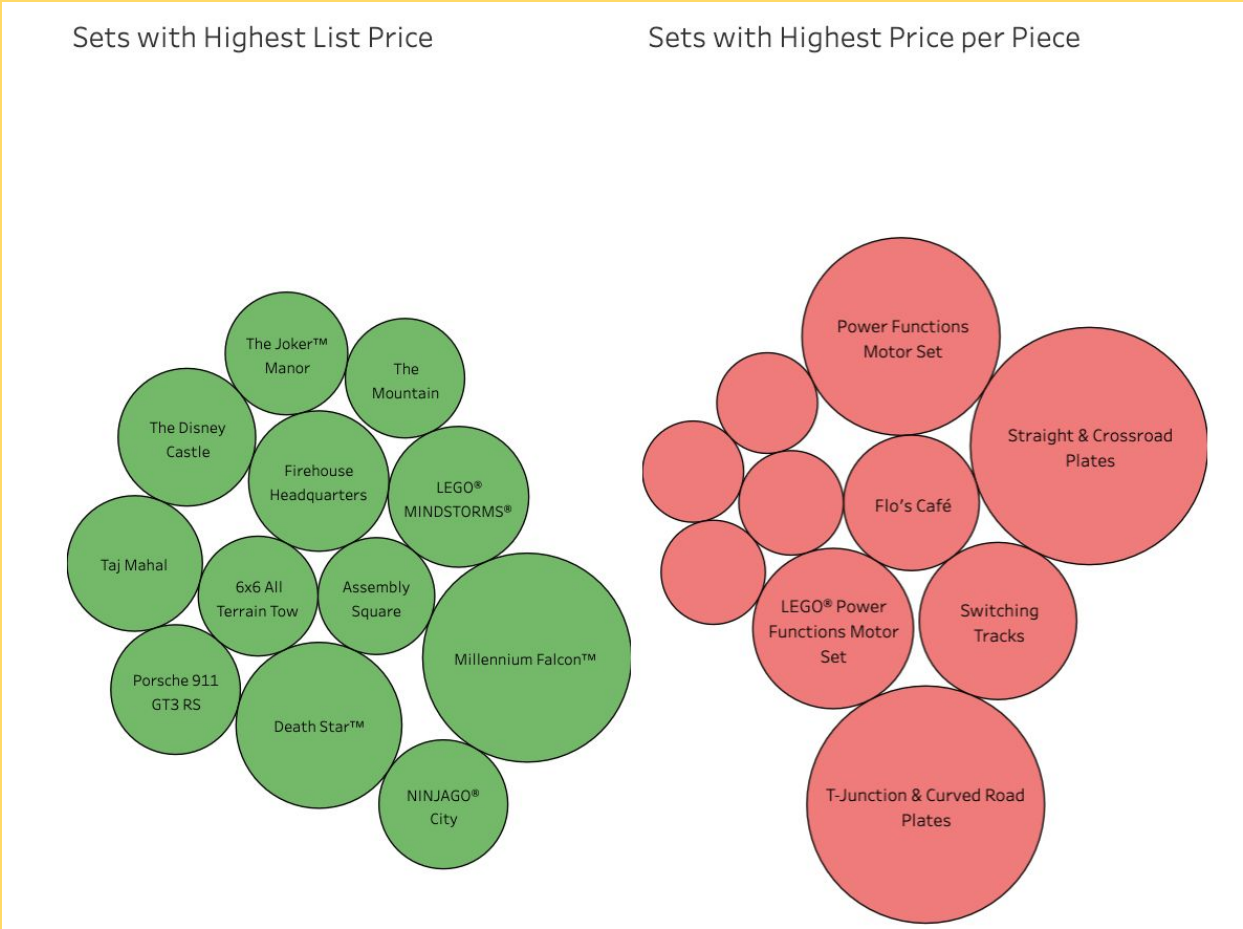
Exploratory Data Analysis



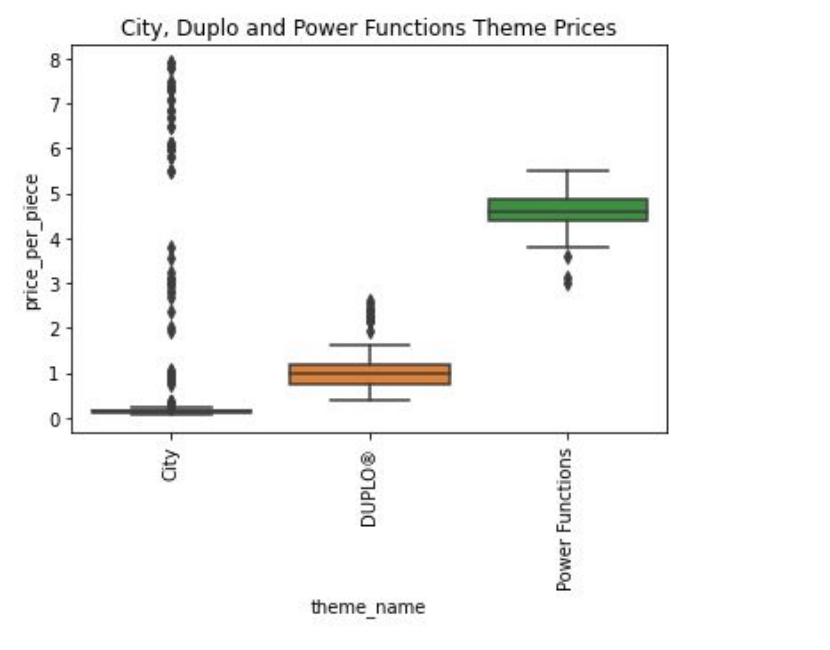
Exploratory Data Analysis



List Price Verssus Price per Piece



Special Cases



Price Prediction Modeling

Data Layout:

1. Price prediction on US and Canada Data
2. Price Prediction on Europe Data

Models Evaluated:

Tree models: Decision Tree, Random Forest Regression, XGBoost Regression, Linear Regression, K Nearest Neighbours.

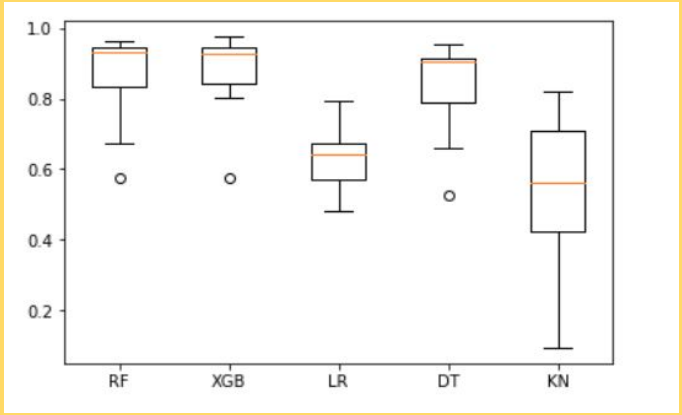
Model Selected:

XGBoost

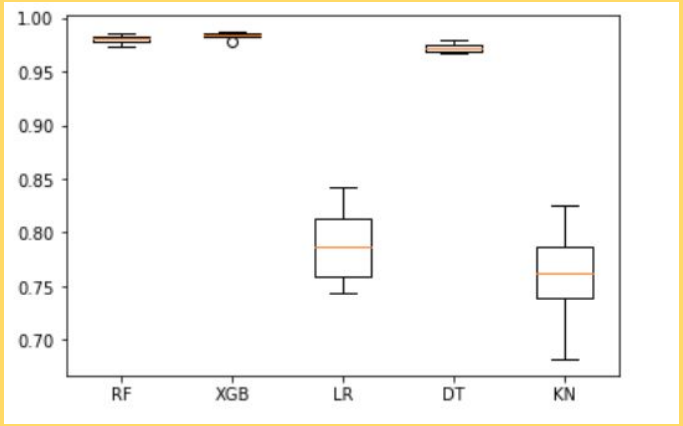
Model Selection and Performance

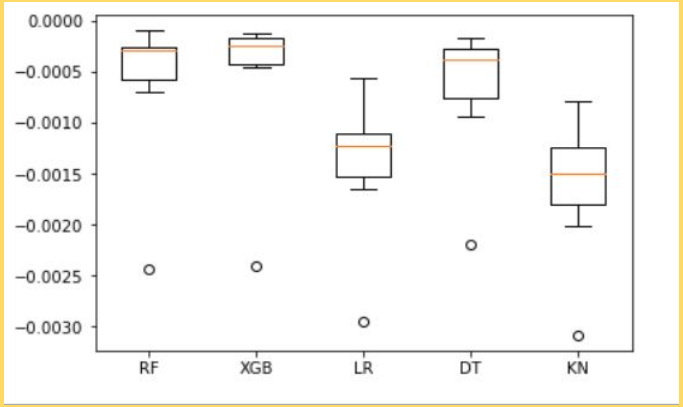
US and Canada Data

Europe, Australia, New Zealand Data

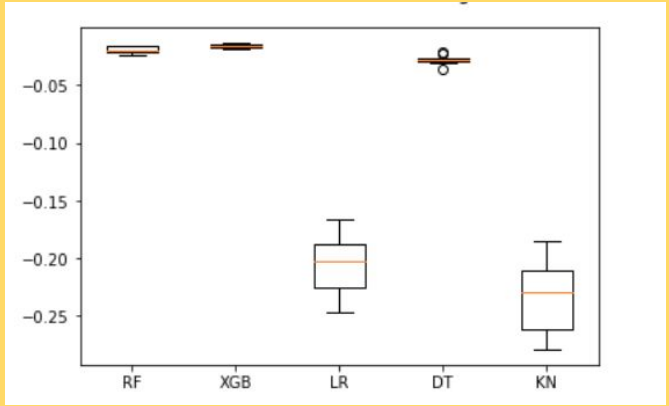


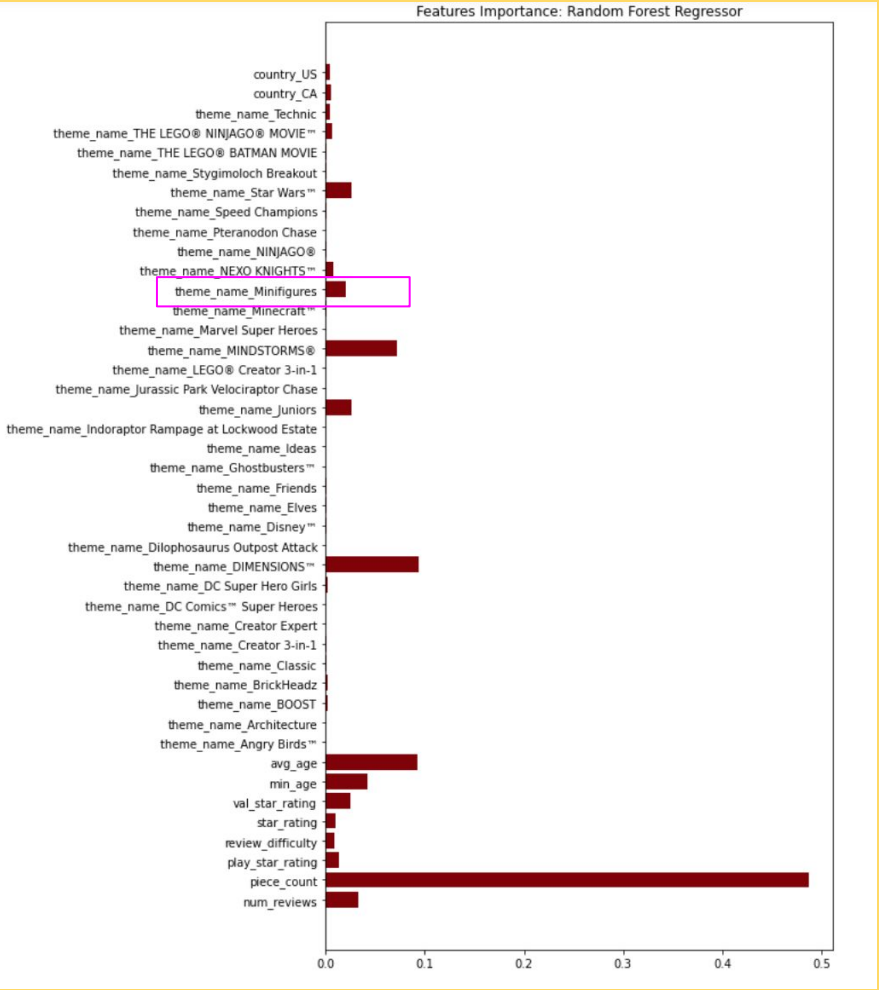
R^2

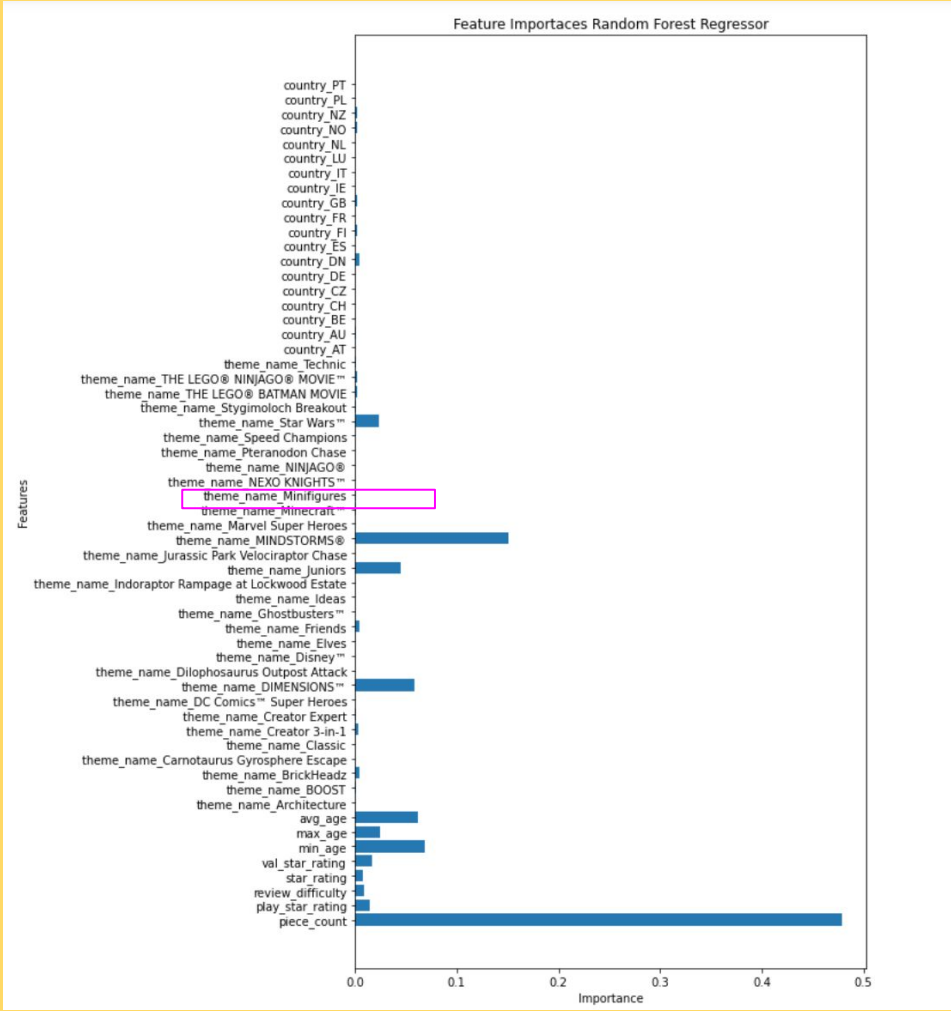




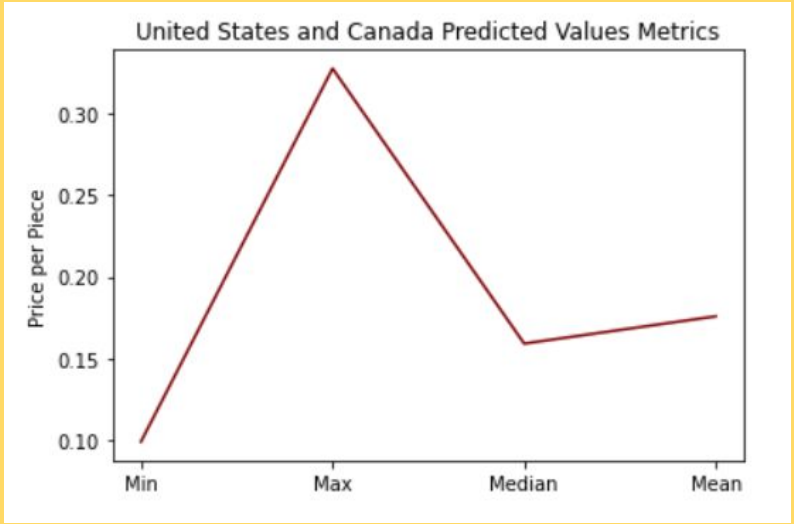
mse

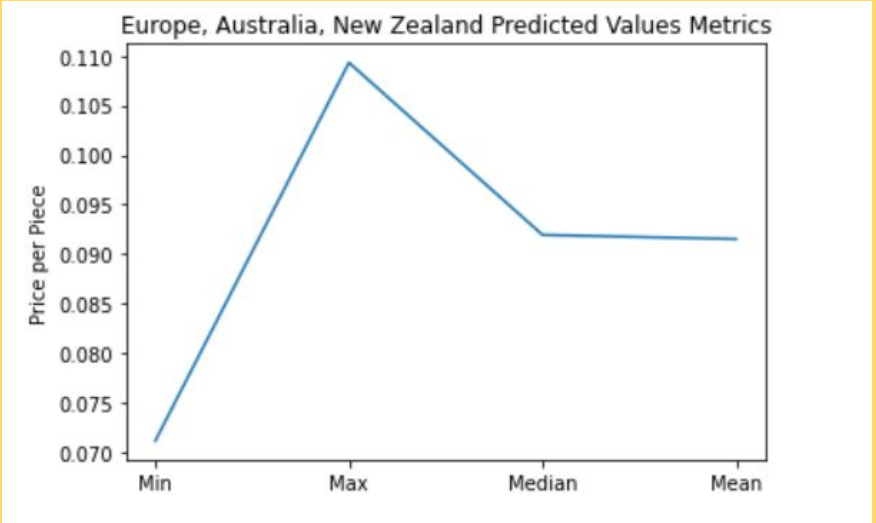



Features Importance: Random Forest Regressor
country_US
country_CA
theme_name_Technic
theme_name_THE LEGO® NINJAGO® MOVIE™
theme_name_THE LEGO® BATMAN MOVIE
theme_name_Stygimoloch Breakout
theme_name_Star Wars™
theme_name_Speed Champions
theme_name_Pteranodon Chase
theme_name_NINJAGO®
theme_name_NEXO KNIGHTS™
theme_name_Minifigures
theme_name_Minecraft™
theme_name_Marvel Super Heroes
theme_name_MINDSTORMS®
theme_name_LEGO® Creator 3-in-1
theme_name_Jurassic Park Velociraptor Chase
theme_name_Juniors
theme_name_Indoraptor Rampage at Lockwood Estate
theme_name_Ideas
theme_name_Ghostbusters™
theme_name_Friends
theme_name_Elves
theme_name_Disney™
theme_name_Dilophosaurus Outpost Attack
theme_name_DIMENSIONS™
theme_name_DC Super Hero Girls
theme_name_DC Comics™ Super Heroes
theme_name_Creator Expert
theme_name_Creator 3-in-1
theme_name_Classic
theme_name_BrickHeadz
theme_name_BOOST
theme_name_Architecture
theme_name_Angry Birds™
avg_age
min_age
val_star_rating
star_rating
review_difficulty
play_star_rating
piece_count
num_reviews
0.0
0.1
0.2
0.3
0.4
0.5

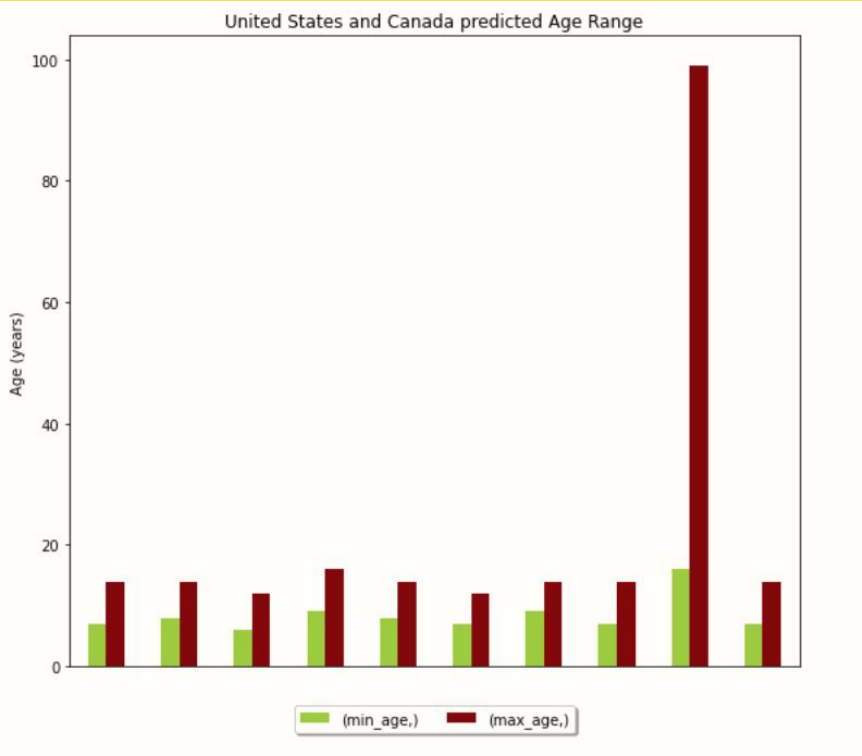

Feature Importaces Random Forest Regressor
Features
country_PT
country_PL
country_NZ
country_NO
country_NL
country_LU
country_IT
country_IE
country_GB
country_FR
country_FI
country_ES
country_DN
country_DE
country_CZ
country_CH
country_BE
country_AU
country_AT
theme_name_Technic
theme_name_THE LEGO® NINJAGO® MOVIE™
theme_name_THE LEGO® BATMAN MOVIE
theme_name_Stygimoloch Breakout
theme_name_Star Wars™
theme_name_Speed Champions
theme_name_Pteranodon Chase
theme_name_NINJAGO®
theme_name_NEXO KNIGHTS™
theme_name_Minifigures
theme_name_Minecraft™
theme_name_Marvel Super Heroes
theme_name_MINDSTORMS®
theme_name_Jurassic Park Velociraptor Chase
theme_name_Juniors
theme_name_Indoraptor Rampage at Lockwood Estate
theme_name_Ideas
theme_name_Ghostbusters™
theme_name_Friends
theme_name_Elves
theme_name_Disney™
theme_name_Dilophosaurus Outpost Attack
theme_name_DIMENSIONS™
theme_name_DC Comics™ Super Heroes
theme_name_Creator Expert
theme_name_Creator 3-in-1
theme_name_Classic
theme_name_Carnotaurus Gyrosphere Escape
theme_name_BrickHeadz
theme_name_BOOST
theme_name_Architecture
avg_age
max_age
min_age
val_star_rating
star_rating
review_difficulty
play_star_rating
piece_count
0.0
0.1
0.2
0.3
0.4
0.5
Importance

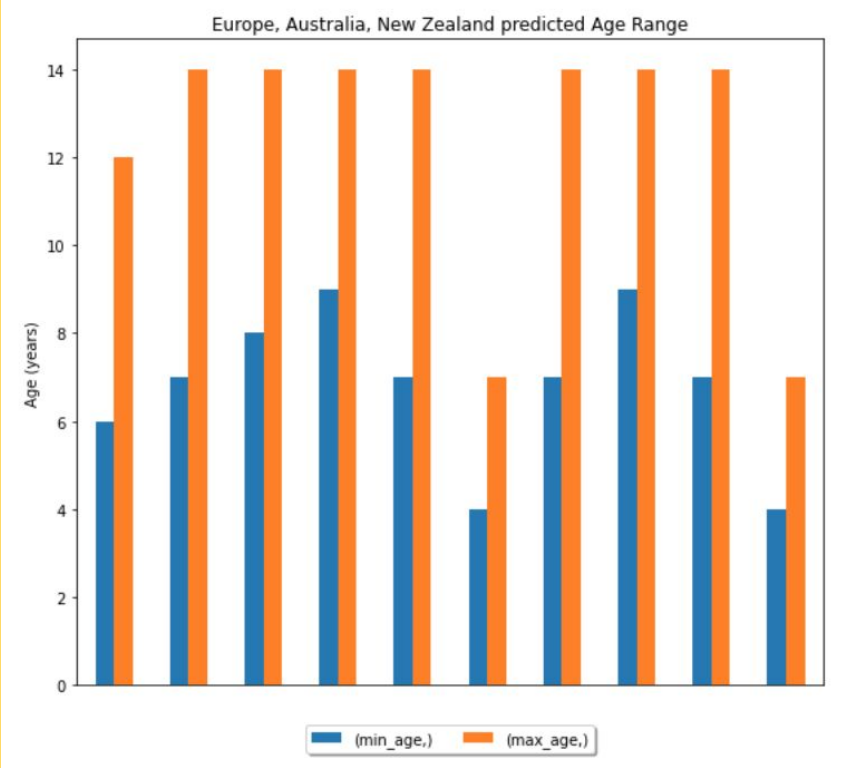
Predicted Data Comparison: Price per Piece



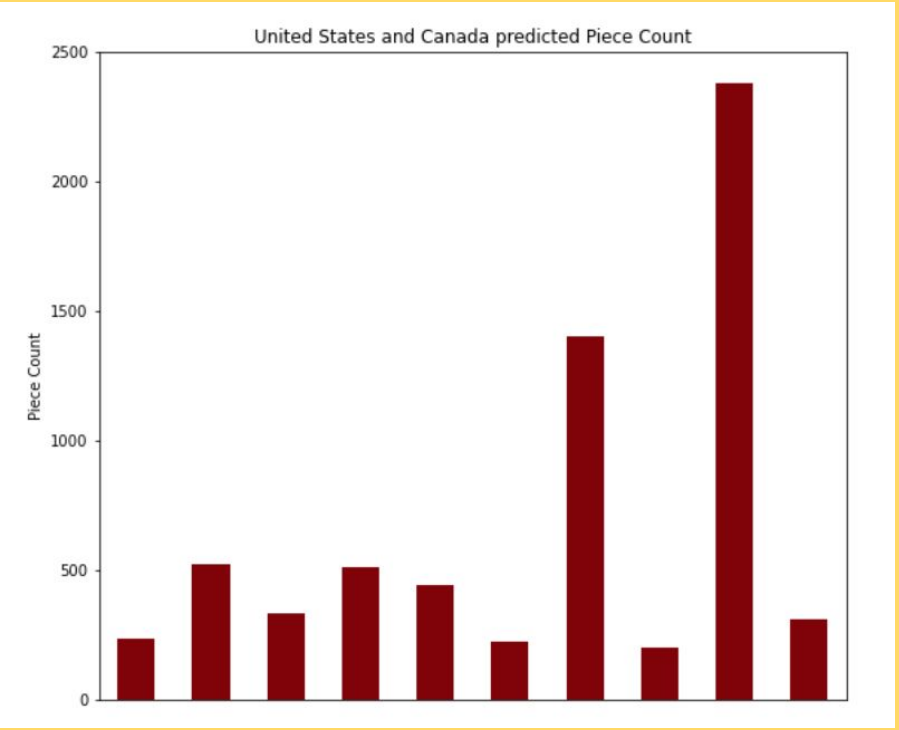


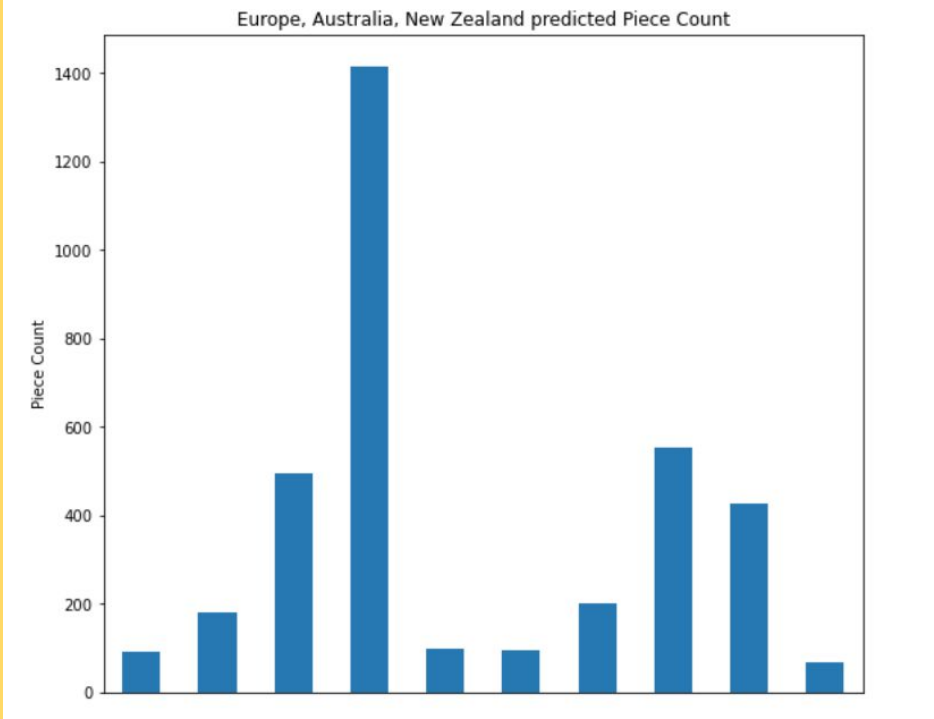
Predicted Data Comparison: Min and Max Age





Predicted Data Comparison: Piece Count





Key Takeaways for Future Business

- Customers in US and Canada are more likely to pay higher price per piece, although, average set price is lower.
- More complicated sets, containing more pieces are more likely to get higher reviews.
- ‘Hit or miss’ with movie character minifigures or sets. Chewbacca and Aquaman minifigures have lowest reviews, while Flo’s Cafe set have splendid reviews despite high price.
- US and Canada have more adult LEGO enthusiasts.

Future Improvements

- More data on Reviews: can predict customer satisfaction and preferences.
- Clear price information (differences in countries about tax inclusion in the price.
- More data on the special Lego Sets (City, Duplo, Power Functions)
- Data in the Asia continent for worldwide analysis

DATA

SORTED

ARRANGED

PRESENTED
VISUALLY

EXPLAINED
WITH A STORY

ACTIONABLE
(USEFUL)