CVRD Mask: Market Analysis (Amazon.com)

By: Julian Murillo
julian@cvrdmask.org

Initial Task:

"Analyze the market for kids face masks. There are a lot of sellers these days. We need to understand what types of masks they are offering and how our masks compare. We would need to compare things like what material are they using, what are their price points, reviews, ratings etc."

Methods:

- Exploratory Analysis: Visualizations
- Summary Statistics
- Ordinary Least Squares (OLS) Regression Analysis

Steps Involved:

1. Search Amazon
2. Scrape Data (Go to a URL and systematically extract data from the website using code)
3. Import to CSV file
4. Clean Data and format in 1NF (1st Normal Form)
5. Begin Analysis

Data Collection:

URL (General Search): https://www.amazon.com/s?k=face+mask+for+virus&qid=1592270249&ref=sr_pg_1

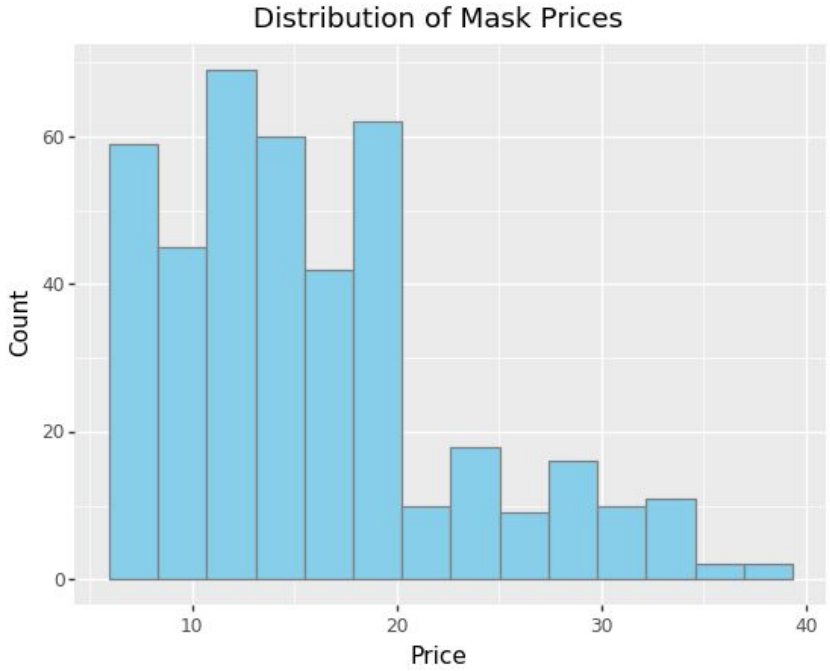
URL (Childrens Mask Search): https://www.amazon.com/s?k=childrens+face+mask&crid=2XM5SAT5P3J0T&sprefix=children%2Caps%2C205&ref=nb_sb_ss_midas-iss_1_8

- Total initial products: 644
- Total Variables: 11 (**1 target: 'product_reviews'**), 4 binary (0 or 1).
- **product_name, product_url, product_price, product_rating, product_shipping, product_reviews, childrens_mask, z_score_price, ear_loops, mask_pack, reusable**

From a Business Perspective:

- Amazon is the largest eCommerce company in the world
- Generating more Amazon Reviews ranks you higher in Amazon's search algorithm.
- Ranking higher generates more user/click traffic and therefore more revenue
- Searching for: key indicators that correlate with reviews
- Decided to try to predict Amazon Reviews using:
    - **product_price, product_rating, product_shipping, childrens_mask, ear_loops, mask_pack, reusable**

Exploring Raw Data: Visualizations

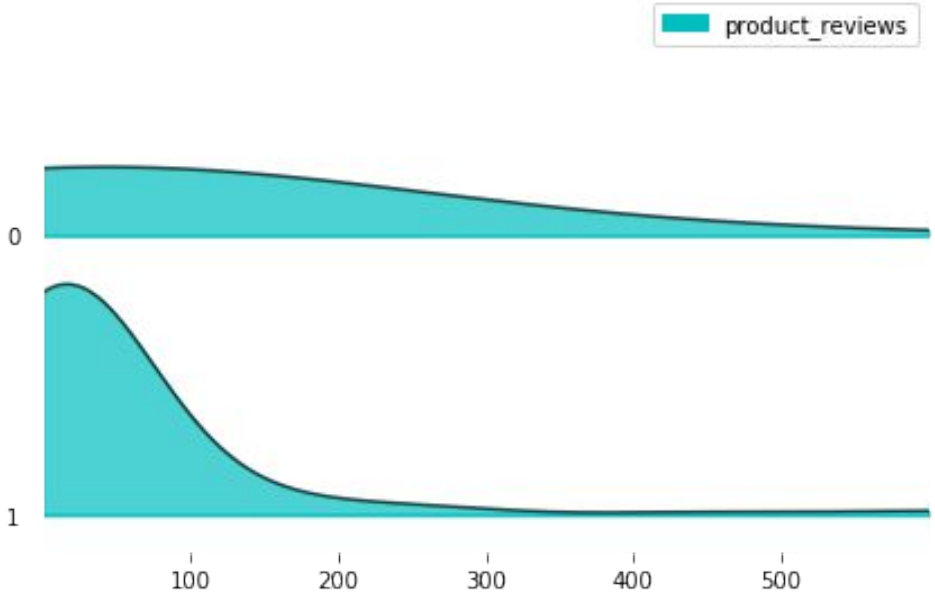


- Skewed right and not normally distributed with a mean price of: $15.79
- Keep prices around this mean to stay competitive in Amazon marketplace

Exploring Raw Data: Visualizations

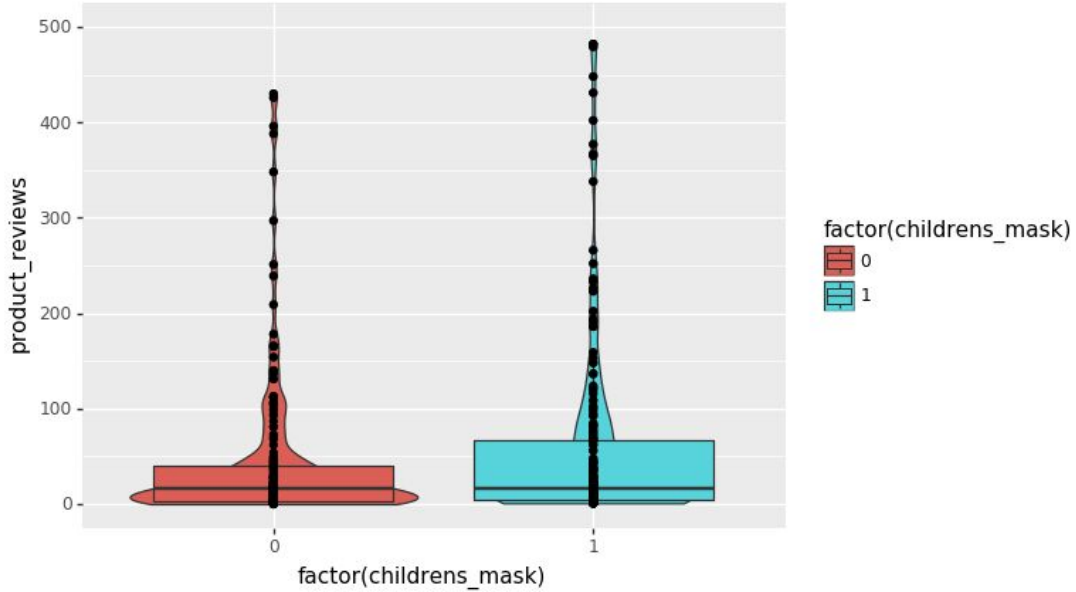
- Price doesn't look to have any effect on rating.
- However: shipments fulfilled by Amazon have much better ratings on average (lower left corner)

Exploring Raw Data: Visualizations



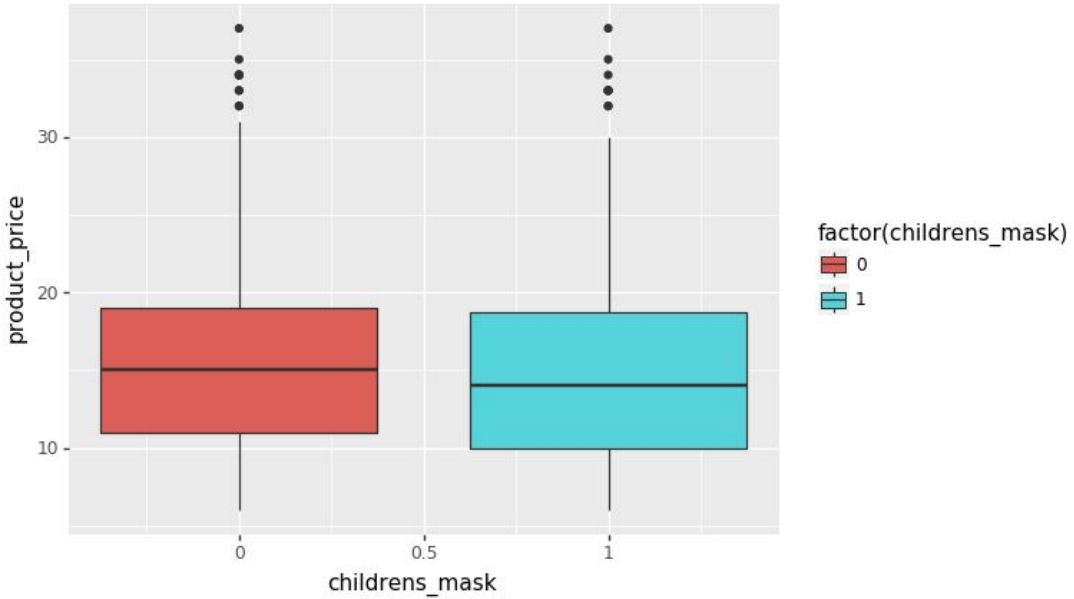
- 0 = Disposable
- 1 = Reusable
- The fill is population density
- The x-axis is the number of product reviews
- Disposable masks seem to have more product reviews

Exploring Raw Data: Visualizations



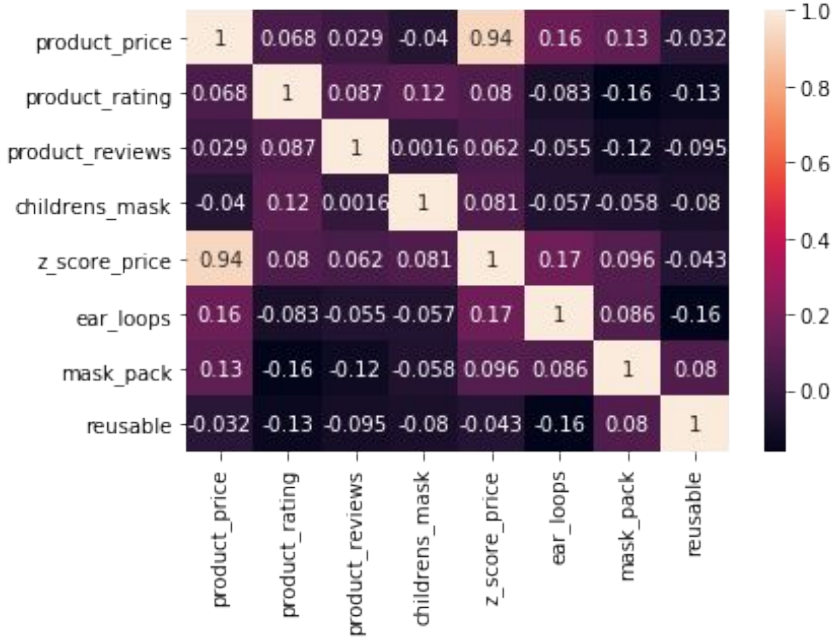
- 0 = Adult Mask
- 1 = Child Mask
- Not a massive difference in the number of reviews of childrens masks vs adult masks.
- But still some difference.
- There is a greater proportion of high-number-of-reviews for childrens masks vs adult masks.

Exploring Raw Data: Visualizations



- 0 = Adult Mask
- 1 = Child Mask
- Not a crazy difference (~ 2 dollars) in price points for Childrens mask
- Which is good, high prices drive profit

Regression Analysis: OLS



- Correlation Matrix to verify that Gauss Markov Assumptions are correct
- They are, no problems with multicollinearity between variables
- Must do this before beginning regression analysis

Quick Note on OLS

- This is the simplest model that someone could make.
- I elected not to use K-Fold Cross Validation, bootstrap aggregation or any more complicated models.
- I was going for parsimony
- I am attempting to predict Amazon Reviews (Target Variable)
-

Regression Analysis: OLS Output

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	73.3810	139.214	0.527	0.598	-200.286	347.048
product_price	3.1725	3.564	0.890	0.374	-3.833	10.178
product_rating	31.6528	30.182	1.049	0.295	-27.678	90.984
childrens_mask	-19.7341	49.665	-0.397	0.691	-117.366	77.898
ear_loops	-95.9198	75.113	-1.277	0.202	-243.577	51.737
mask_pack	-103.9034	51.081	-2.034	0.043	-204.319	-3.488
reusable	-92.5366	51.148	-1.809	0.071	-193.084	8.011

- Interestingly, the two things that seems to drive the number of reviews a product gets are the price and (obviously) the product rating.
- If you increase the product price by 1 dollar, you will expect to see around 3 more reviews on average. Similarly, if you increase the product rating by 1 star, you can expect to see an increase of about 31 reviews on average. (look at ‘coef’ column)

Model Evaluation: Robustness Checks

```
# Very high mean squared error lol

mean_squared_error(Y, reviews_pred)
```

```
242817.53264493175
```

```
r2_score(Y,reviews_pred)
```

```
0.029941815894697 02
```

Very Small R2
The R2 tells us the % of variation in Amazon Reviews that is explained by all our predictor variables. About 3% in this case :/

OLS Regression Results

Dep. Variable:	product_reviews	**R-squared:**	0.030
Model:	OLS	**Adj. R-squared:**	0.016
Method:	Least Squares	**F-statistic:**	2.099
Date:	Fri, 19 Jun 2020	**Prob (F-statistic):**	0.0524
Time:	11:39:18	**Log-Likelihood:**	-3161.9
No. Observations:	415	**AIC:**	6338.

Market Analysis Conclusion:

Childrens masks, ear-loops, packs of masks, and reusable masks were inversely (negatively) correlated with the number of reviews a product had. Refer to Coefficients list.

Overall, these variables combined account for about 3% of the variation in number of reviews... not a very solid estimate (pretty bad actually). We would need much more information/many more variables to generate a predictive model to really see what drives Amazon reviews.

Market Analysis Conclusion:

- The higher the price is on your product, the more likely it is to get reviews
- Also if your product has more positive ratings, then it will tend to generate more product reviews.
- Children's masks, Earloops, and reusable material generally don't increase the number of reviews your product gets.