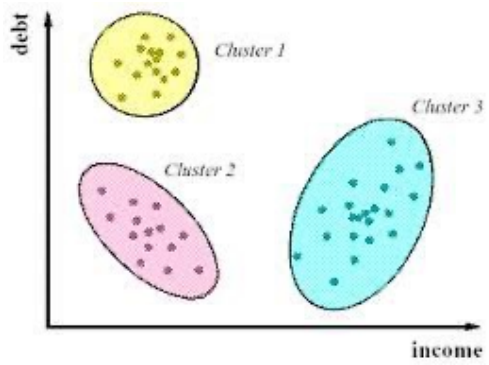
Nordstrom Skin Care Products – Final Project

By Aqsa Majeed, Ava Prativadhi, Sharon Corrales

RECAP

- The goal is to find a relationship between price and high rated skin care products.
- This was to help consumers find affordable and effective skin care products.

Our dataset consist of skincare product brands with different ratings, price, and number of rates. Throughout our time working with the dataset, we were able to investigate and find different relationships between each of the main variables within the dataset. We will discuss this and the correlations we found which helped us deeply understand whether there is a present relationship between price and the skin care products given in the dataset.




```
set.seed(88)
cluster.skincare <- kmeans(skincare[,2,4], 5, nstart = 20)
cluster.skincare
```

```
K-means clustering with 5 clusters of sizes 4, 121, 242, 383, 91

Cluster means:
       [,1]
1 403.44000
2  60.47493
3  32.82958
4  16.65721
5  96.65255
```

Modeling (Hyperparameter Tuning)

- Unsupervised learning methods: KMeans Clustering
- **random_state :** random seed of 88 .
- **N_clusters:** 5 clusters
- **Init:** Initial cluster centroids (5)
- **nstart:** multiple initial configurations and reports the best.

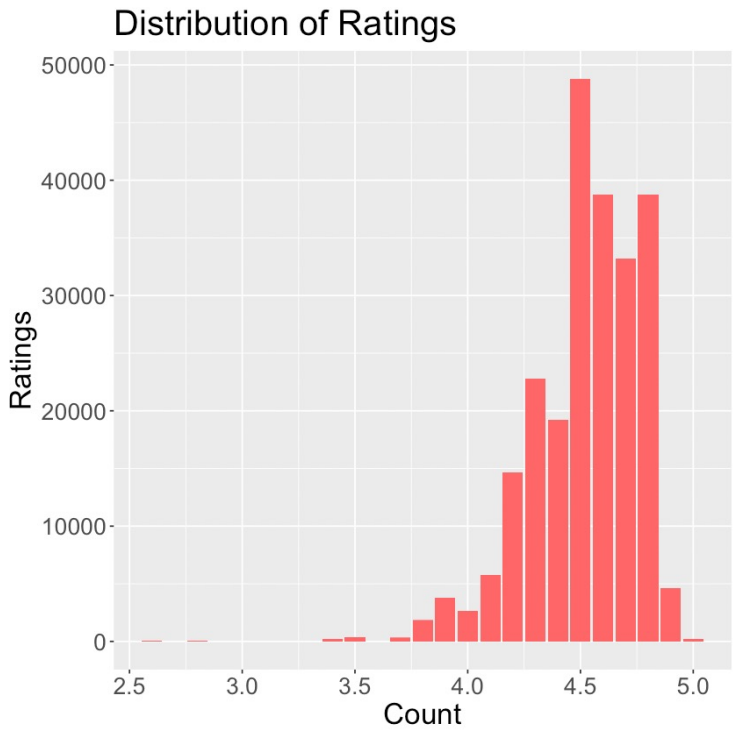
Model (testing) slide 1

```
K-means clustering with 5 clusters of sizes 4, 121, 242, 383, 91

Cluster means:
       [,1]
1 403.44000
2  60.47493
3  32.82958
4  16.65721
5  96.65255


Within cluster sum of squares by cluster:
[1] 56082.495  8731.370  8644.788  6728.729 24349.186
 (between_SS / total_SS =  91.2 %)
```

- Created 5 clusters
- 91.2 # total variance
- The k-means clusters have homogeneity: all the observations with the same class label are in the same cluster

Model (testing) slide 2

- Most of the data falls between 4 and 5.
- Built a model on the training data using regression analysis to check for the trend
- Did a k-fold Cross validation with 5 folds




```
Call:
lm(formula = stars ~ price, data = skincare, na.action = na.exclude)

Residuals:
     Min       1Q   Median       3Q      Max
-1.90344 -0.17844  0.09575  0.20156  0.51866
```

```
Residual standard error: 37.67 on 835 degrees of freedom
Multiple R-squared:  0.007556,  Adjusted R-squared:  0.001613
F-statistic: 1.271 on 5 and 835 DF,  p-value: 0.2741
```

Model (output) slide 1

- **Section 1:** Installing four packages and the library functions for the four packages. The four packages used were dplyr, ggplot2, psych, stats, car
- **Section 2:** The first 10 rows of the dataset, the basic statistic for the data variables, the correlation between the two variables. The two variables are “stars vs price” and “ratings vs stars”.

```
describe(skincare)
```

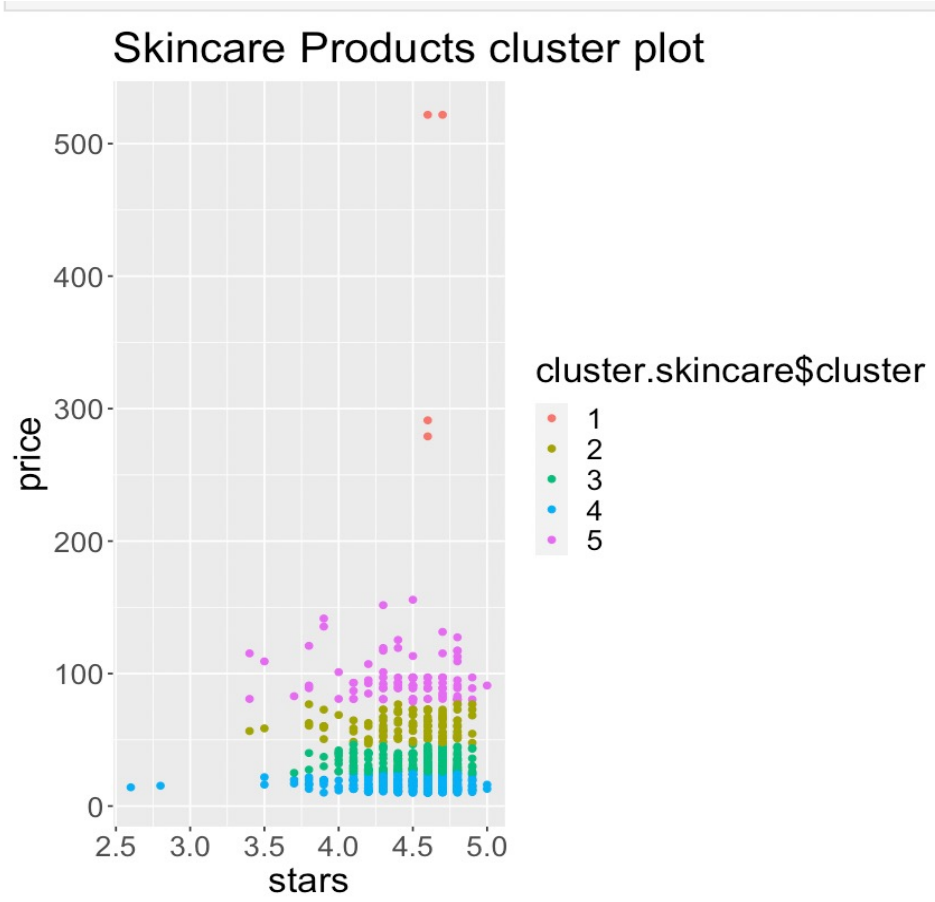
A psych: 5 × 13

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
X	1	841	1528.513674	1030.9573383	1419.000	1511.554235	1439.60460	0.00	3243.000	3243.000	0.068567288	-1.463263	35.55025305
price	2	841	38.110678	37.6989041	26.292	32.060266	17.39238	10.11	521.771	511.661	6.152413661	66.950076	1.29996221
title*	3	841	415.969084	239.5391687	417.000	415.961367	306.89820	1.00	831.000	830.000	-0.002509705	-1.202257	8.25997133
stars	4	841	4.496552	0.2910163	4.600	4.529569	0.29652	2.60	5.000	2.400	-1.404557803	3.830322	0.01003504
vote	5	841	280.752675	734.4843482	129.000	168.716196	93.40380	51.00	13221.000	13170.000	13.583601543	227.928931	25.32704649

The outputs products for section 1 & 2

```
cor(skincare$stars, skincare$price)
```

-0.0372493710558717

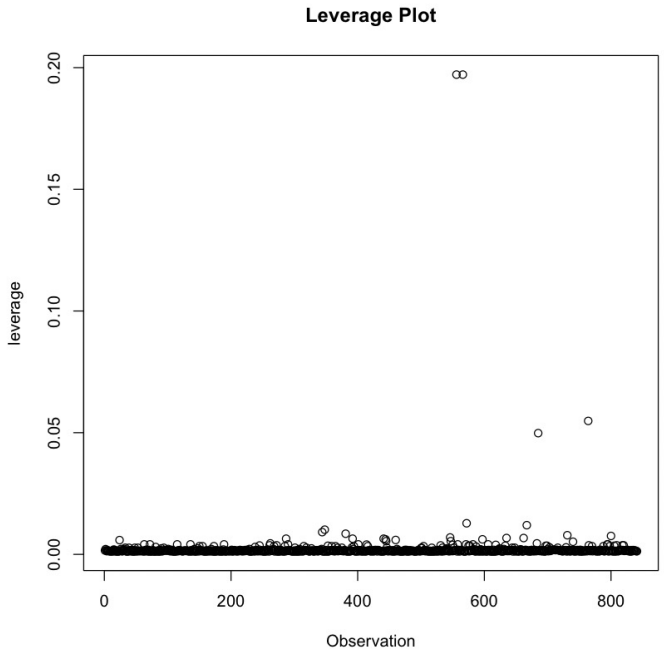


Model (outputs) slide 2

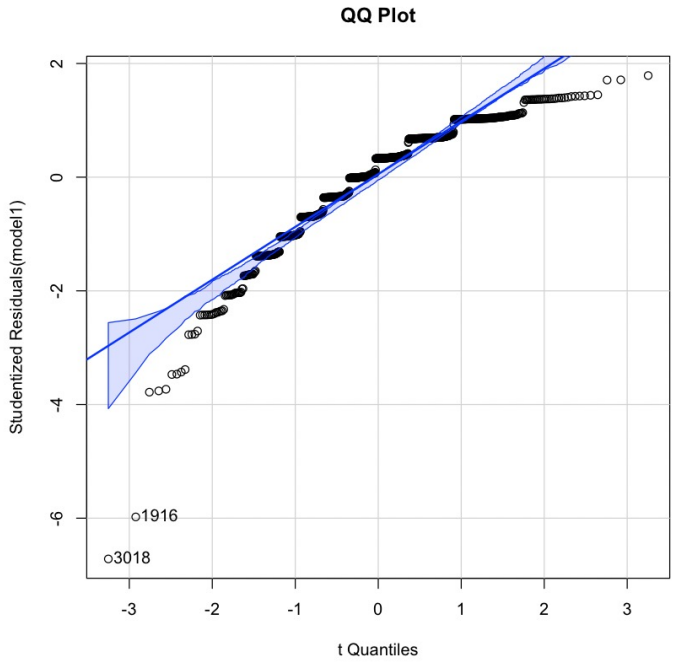
- **Section 3:** K means clustering with 5 clusters along with the clustering vector and a plot for skincare products cluster plot
- **Section 4:** It includes the residuals, coefficients, and residual standard error. GGplot with price of product vs rating received, QQ plot, leverage plot, and Cook's distance.

The outputs for sections 3 & 4

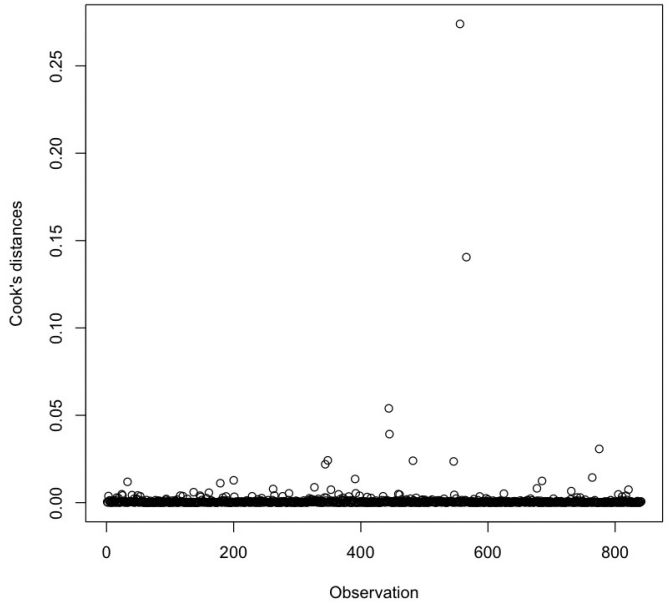
- Leverage Plot
- QQ Plot
- Cook's Distance



The points had a low leverage so the influence on the model parameters was small.

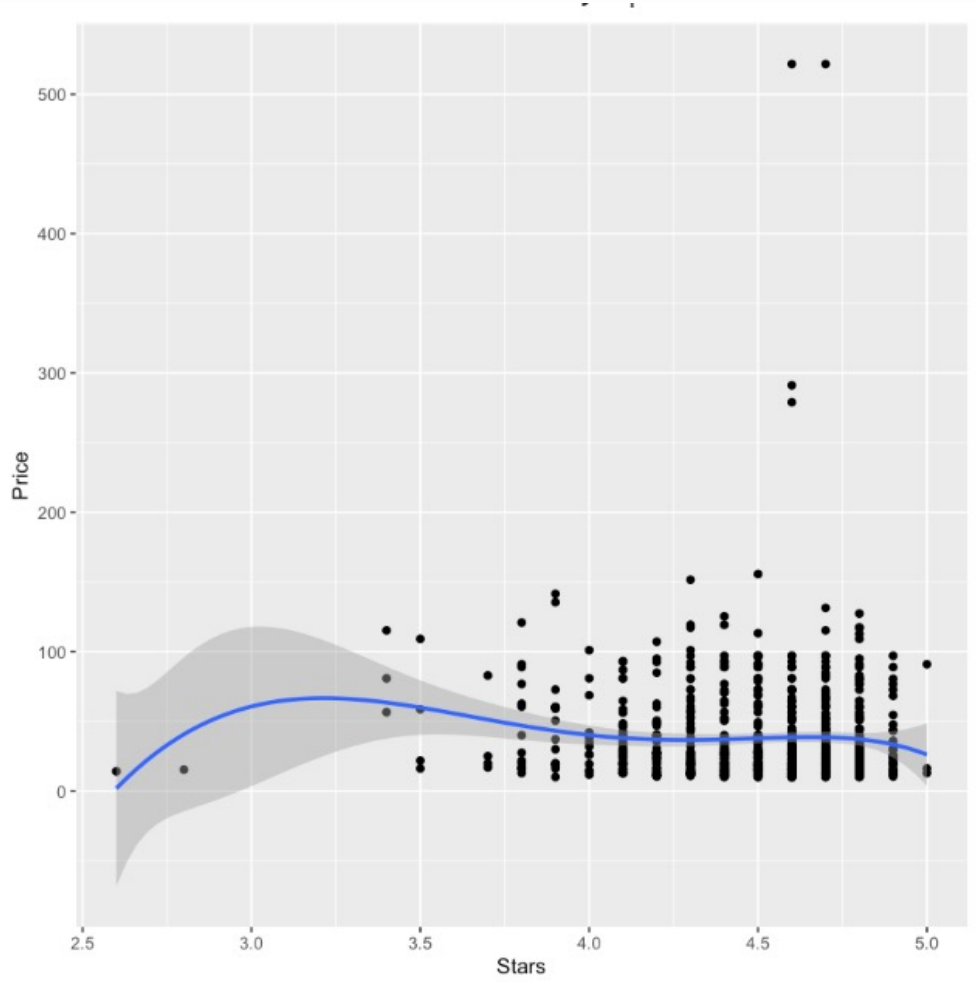


The lines didnt fall in the linear manners but has a curve which means there are outliers.



The threshold is at approximately at 0.10 with two outliers.

Final Model Product



Modeling (assumptions and limitation)

Assumption 1: We assume that the products with higher rating are more expensive.

Assumption 2: High rated products that are more expensive are made with quality ingredients.

Assumption 3: We assumed our plot graphs would look more scattered.

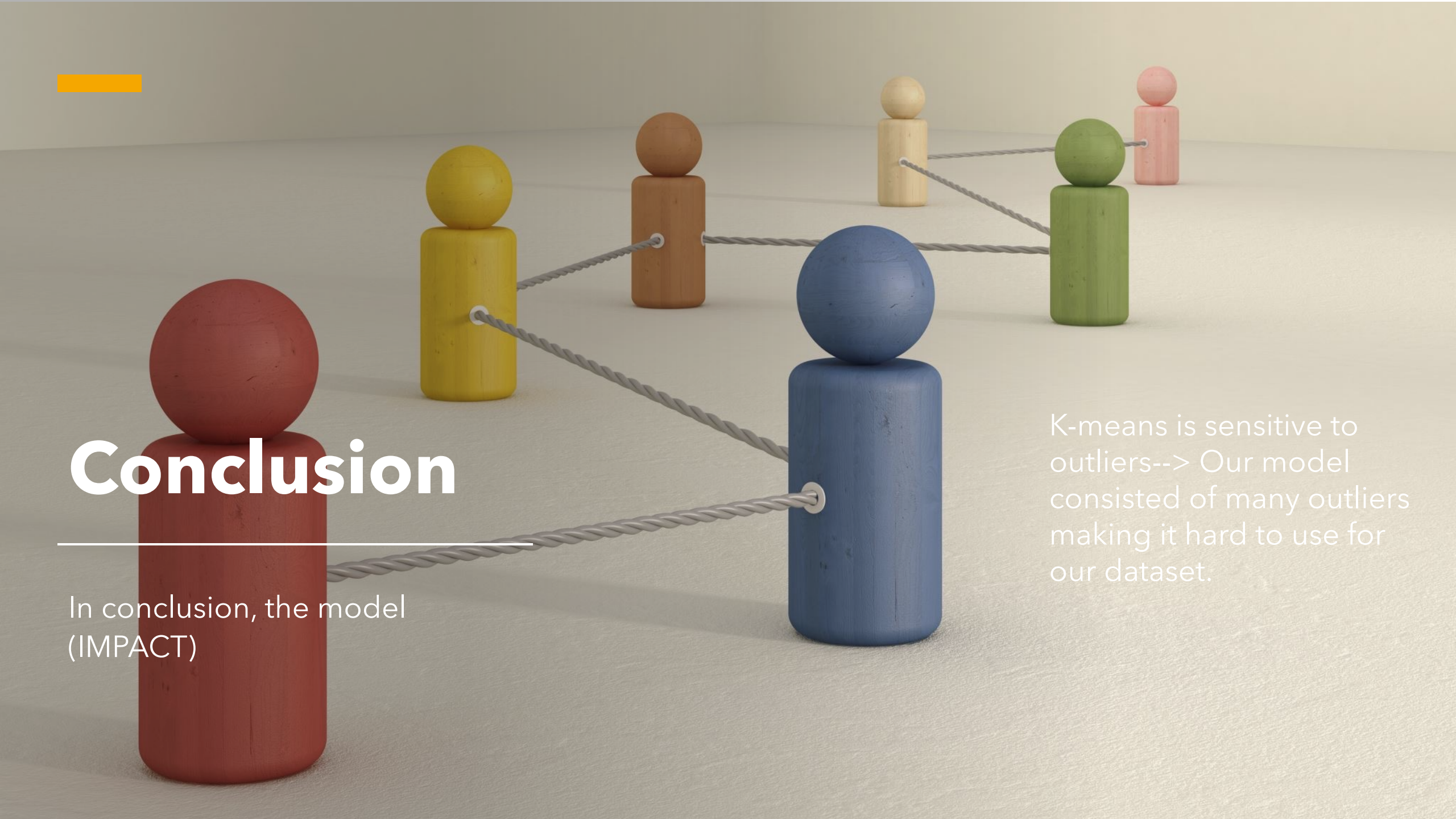
Limitation 1: We are limited to the information that the dataset provides us. So, we were not be able to analyze the different ingredients and their relationship to the price.

Limitation 2: The correlation between the variables is very low and the graphs didn't come out the way we expected them to.

Limitation 3: To keep the graph simple we only used two variables and couldn't use more.

Evaluation

- The model reflected a constant relationship between price and rating.
- This means there is no relationship between the price and ratings of the products within the dataset.
- The dataset consisted a lack of products with different rating (3-4 stars only!). The pricing was very difficult to read.



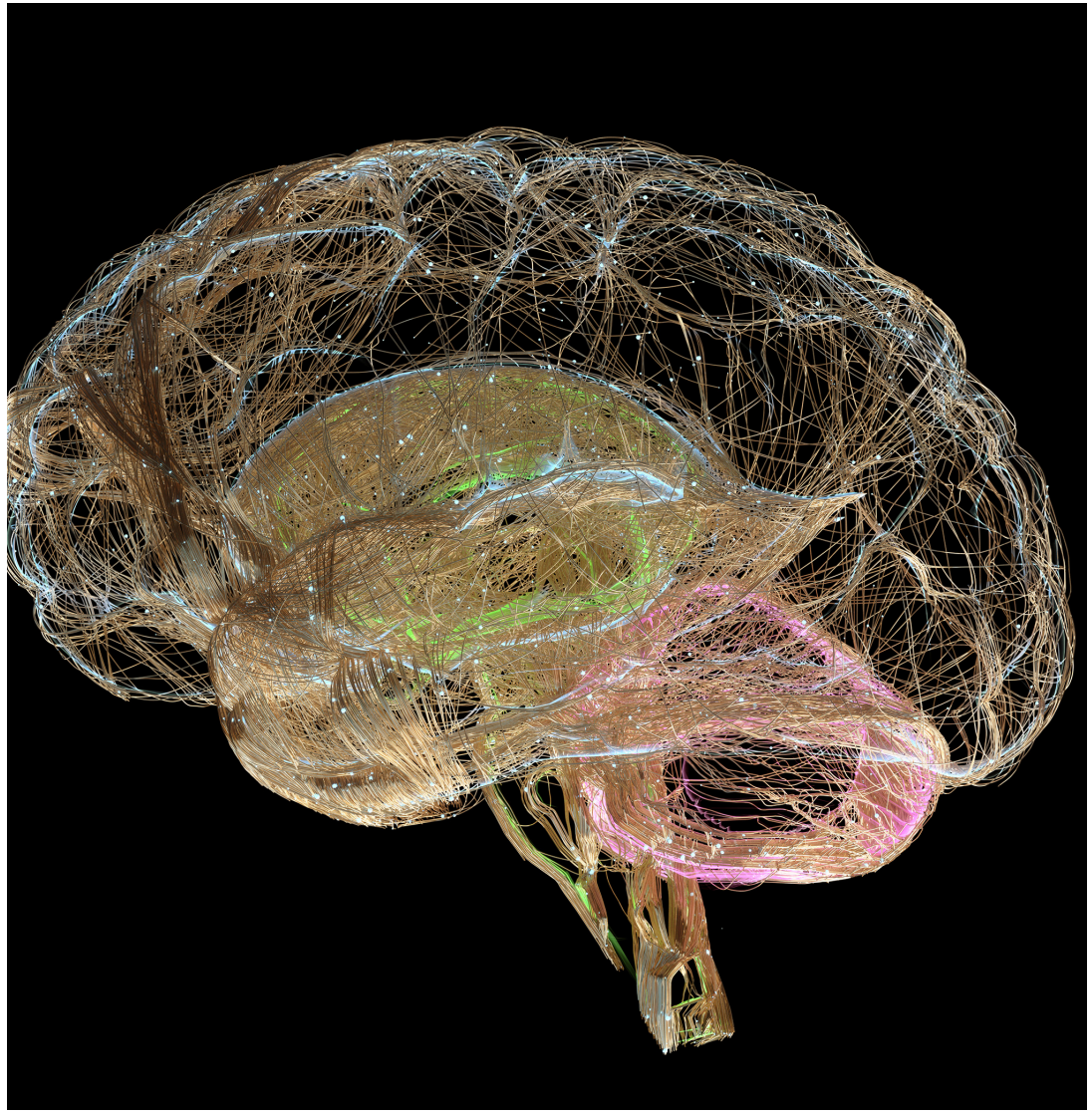
Conclusion

In conclusion, the model (IMPACT)

K-means is sensitive to outliers--> Our model consisted of many outliers making it hard to use for our dataset.

Things we wanted to do but couldn’t

- Use more variables along with prices and ratings.
- Use more than one dataset for our model.
- Use 10 folds instead of 5 folds.
- Use a dataset of other than Nordstrom products.
- make a graph that shows the different skincare products and their ratings to see which product was rated the highest and which one had the lowest rating.
- Due to our dataset only including high end products it would’ve been interesting to see the ratings of both in one dataset.



References

- *Learn by marketing*. Learn by Marketing | Data Mining + Marketing in Plain English. (n.d.). Retrieved July 26, 2022, from https://www.learnbymarketing.com/methods/k-means-clustering/
- Khan, M. (2017, August 2). *KMEANS clustering for classification*. Medium. Retrieved July 26, 2022, from https://towardsdatascience.com/kmeans-clustering-for-classification-74b992405d0a
- Team, T. V. (2021, July 6). *Cluster analysis in R - complete guide on clustering in R*. TechVidvan. Retrieved July 26, 2022, from https://techvidvan.com/tutorials/cluster-analysis-in-r/