



Top 500 SKUs - Queen Bed Regression Analysis

Drew Laird

Exclusive Brands

Manager: Meghan Steeves

Objective

To hone in on **what factors (variables)** are the most important in **determining the success of a SKU on Wayfair** using a **multiple linear regression analysis**. Specifically, I'll be looking at the **top 500 queen beds** to narrow down my results and hopefully develop a method that could be quickly applied to other classes. Success, in this case, is going to be measured by **GRS**. Within this domain, I will also focus on to what degree being in a FSB or header brand affects performance to demonstrate a measurable **benefit to being in an exclusive brand**. The results of this project will hopefully have some repeatability, but it is not expected that this model will be overly robust or conclusive because there are a lot of factors at play and tests (for things like collinearity or heteroskedasticity) that will be neglected for the sake of efficiency and simplicity.

Variables

Dependent Variable(s):

- GRS
- *(optional)* Continuous Winner algorithm score

Numerical Independent Variables:

- Price (taken from the month of December)
- Number of reviews
- Average review rating
- Visits
- DWIS%
- Damage rate
- Return rate
- Exclusivity %

*use Brand HUD

Categorical (Dummy) Independent Variables:

- Material
 - Wood, upholstered, metal
- In FSB
- In Header
- Bed type
 - Wingback, Panel, Slat, Four Poster, Murphy, Canopy, Open-Frame
- Storage Included
- Box Spring Requirement/Platform Bed
- Assembly requirement
 - Assembly Required or Partial

Sub-Hypothesis

Given the numerical and categorical variables outlined above, membership in an exclusive brand will lead to increased performance as represented by a positive coefficient for the FSB and Header brand dummy variables.

Data

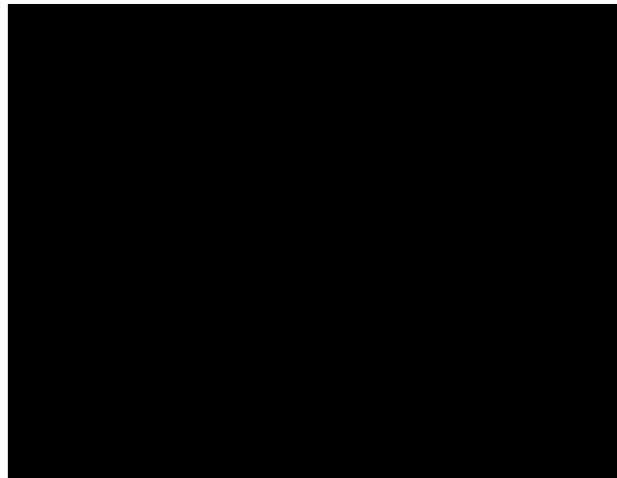
The first problem in gathering the data was finding out which of the SKUs in the bed class were queen beds, and, of those queen beds, which ones were producing the most GRS. To sort for queen beds, I had to pull all the SKUs that have the attribute option “Queen” under the attribute “Bed Size.” This was done through a [script from Sarah Yip](#), an analyst that helped me in the analytics-forum. Once I figured out the class ID for beds (12), and the attribute option ID for queen beds (127), I pulled a google sheet with all the SKU numbers for queen beds.

Next, I used [one of the vetted scripts](#) that pulls SKU-level GRS in the last twelve months. For my data set, I used the data from the entire year of 2021. I then sorted this for every bed SKU. From there, with help of Melissa, I used vlookup to combine these sheets and match the GRS for the list of Queen beds I just gathered. When I sorted by GRS, it gave me the ranked queen bed SKUs by GRS. From there,

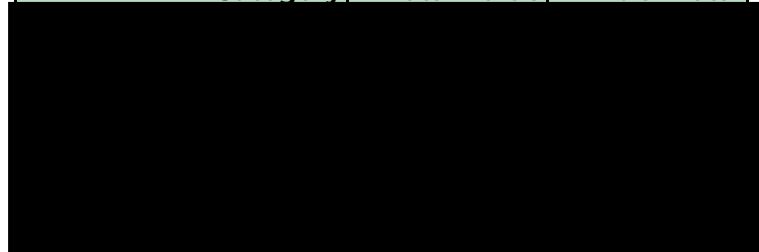
to shrink my sample size to something more manageable, I only considered the top

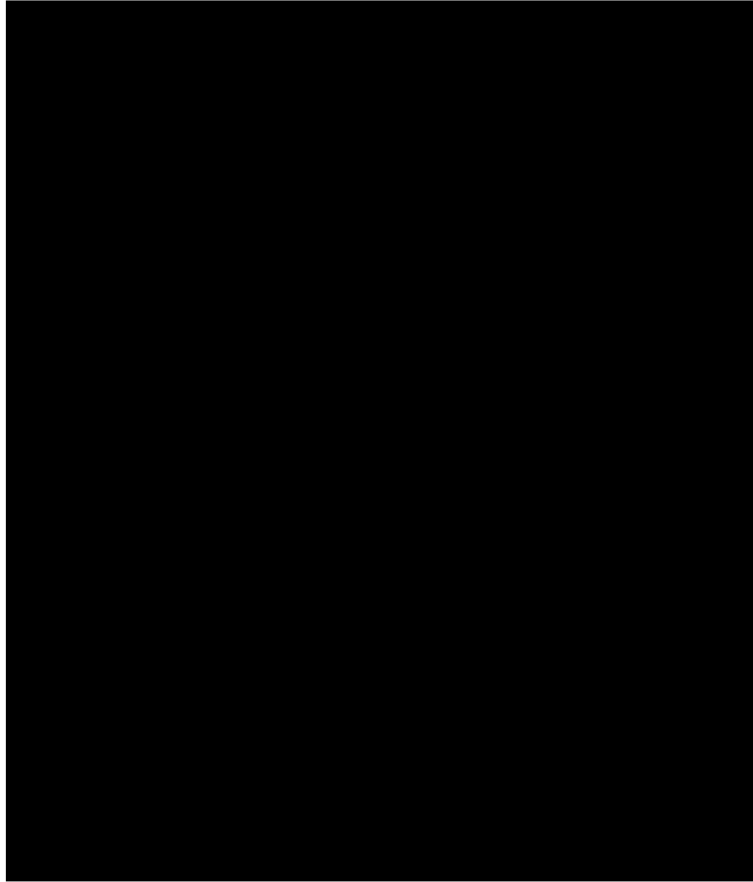
Once I had this list of SKUs, I pulled the Brand HUD to pull most of the numerical data I needed. The harder part was getting the attribute data. Using my list of SKUs, I created another sheet with all the attribute data with [this script](#). After parsing through it, I identified important factors which became the categorical dependent variables above. Then, with help from Christina, I converted all the attributes into 0s and 1s. The 1 signifies true and the 0 signifies false. For example, if a SKU had a value of 1 for Wood, that meant the attribute indicated the bed was made of wood. In statistics, these types of variables are called dummy variables.

Finally, with my finished data set, I took a look at which SKUs were in exclusive brands, and which weren't. You can see a summary of that data in the table below. About half of the assortment was in exclusive brands, which is greater than the total assortment ratio of EB to non-EB.



In the table below, you can see the distribution of the categorical variables in the top 500.

Category	Total Value	% of Total
		



Analysis

After gathering the data, I needed to run a multiple regression on the data. If you don't know what multiple regression means, I'll include a crash course in my presentation. I could not find openly available tools that were approved by Wayfair to run the regression on my full data set. I tried to use Big Query to train a linear machine learning model, but I did not have enough time to complete it and troubleshoot the issues I was getting. There is a project within "[the-fairway](#)" called "[the-fairway](#)" where you can find my data table. I tried running a regression analysis according to "[the-fairway](#)". If you're up to it, you're more than welcome to try to figure that out.

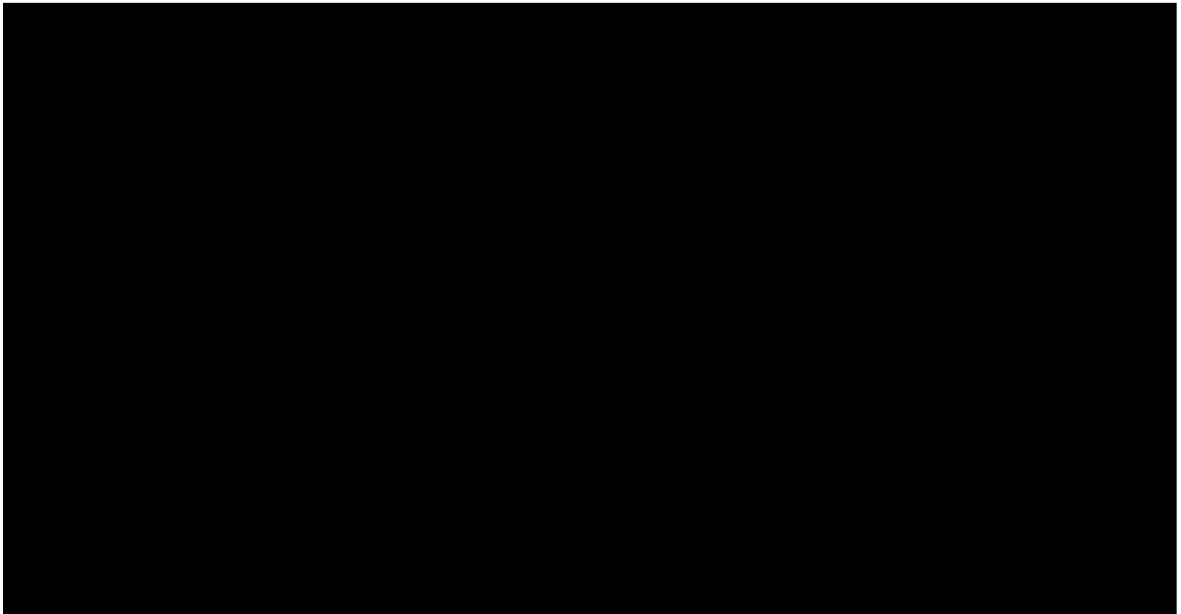
Hard decisions...

Excel offers a free multiple regression tool, and, thanks to Connor, I was able to run this on his computer. [REDACTED] The problem with this regression tool, however, is that it cannot handle any more than sixteen variables. As a result, I was forced to select which sixteen variables I considered the most important. To somewhat get around this, I ran the regression three times, each time with a different combination of variables. The first combination I chose based off of my intuition. In the second sheet in my file, "Possibly Irrelevant Variables," you can see which variables I thought were the most likely to be irrelevant to the model—or statistically insignificant. Here's my reasoning for the highlighted variables:

- **WSCNR** - It seems fairly obvious that revenue is going to be extremely correlated with GRS, WSCNR directly contributes to revenue in the absence of Wayfair's margin.
- **Stars** - There really isn't enough variability there, none of the top SKUs are going to have a rating below four stars or that far from five stars.
- **Damage** - It's probably more likely that returns are going to directly affect GRS rather than damage. Damage itself is probably already correlated with return rate.
- **FSB and Header** - To shrink the number of variables, I combined these two columns into just EB. Either way, we just want to know whether being in an exclusive brand—which has boosting—is increasing GRS.

- [REDACTED]
- [REDACTED]
- [REDACTED]

-
-
-
-



Notable regression takeaways

When you run a multiple linear regression, two things matter more than anything else: the coefficients (especially whether they're positive or negative), and the p-values.

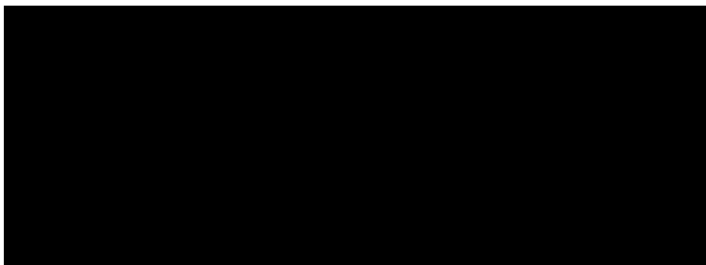
Essentially, the coefficients indicate **how much a singular change in the independent variable will affect GRS**. For example, if the coefficient for price were 100, this means that for every dollar increase in price, GRS would increase by 100. If it were -100, GRS would decrease by 100 for every increase in price. For a dummy variable, let's say metal, if the coefficient is 100, then this means that when the bed is made of metal the independent variable value will be equal to 1. Thus, 1 (the independent variable value) * 100 (the coefficient) = 100, so GRS would increase by 100. When it's not metal, it's $0 * 100 = 0$ so GRS would not change. In effect, the coefficient measures the weight of a difference that changes in the independent variable make. I'll give concrete and visual examples in my presentation or you could research if you need further clarification.

As much as these coefficients matter, they would mean nothing if they weren't statistically significant. Statistical significance, on a high level, just indicates the level of certainty you have that the coefficient your regression gave you wasn't a result of randomness. This is indicated by p-values, this is a measure of probability, so it ranges between 0 and 1. The more statistically significant a p-value is, the closer it

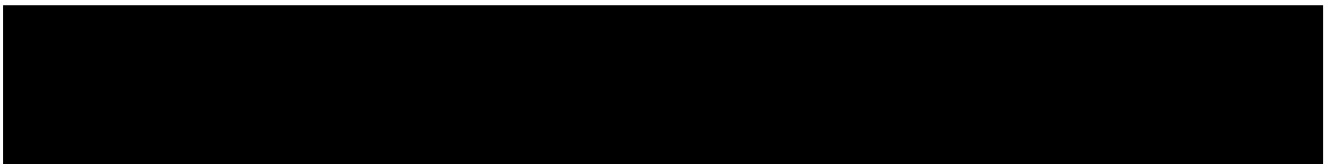
is to zero. Essentially, if you have a p-value of .01, this means the probability that the coefficient you received was due to randomness was .01, or 1%. On the other hand, if it's .9, this means there is a 90% chance that your coefficient value was a result of randomness. In general, the rule of thumb among statisticians is that anything with a p-value less than a .05 is statistically significant.

Because I had to run multiple, mildly different regressions, I focused on just the commonalities between these regressions. Thus, which coefficients were consistently statistically significant, and what their values were. Generally I observed the following variables were the most significant:

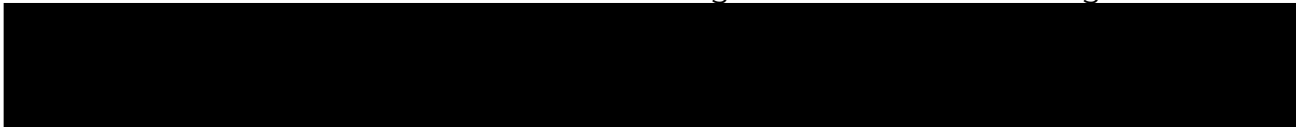
-
-
-
-
-



This variable had a positive coefficient around \$180,000 with a statistical



The variable reviews has the second highest level of statistical significance. In



[REDACTED]

[REDACTED]

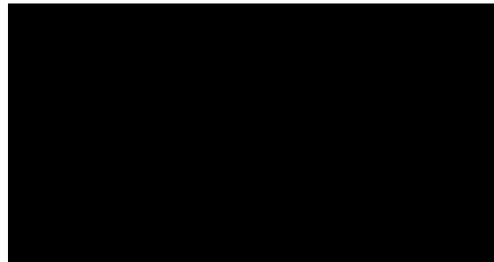
[REDACTED]

[REDACTED]

[REDACTED]



Furthermore, interestingly enough, if we look at another metric, **continuous winner score**, that **paints a similar story**. Courtesy of Eric Guo, I have the top 500 queen bed SKUs CW score, and I compared that for EB and non-EB SKUs. This is a much more robust model that theoretically makes the most accurate predictions on SKU success that Wayfair currently has to offer. The results are as follows:



Either way, due to the imperfections of this model, more exploration is needed to understand the meaning behind these results.

Conclusion

Due to the nature of this model, I can only make a limited set of conclusions. As with many statistical studies, the results leave just as many (or more) questions than answers. While a statistically significant decrease in queen bed revenue for EBs is an interesting find, there are a couple of known flaws with this model that I'll dive deeper in.

Known flaws

- **Not testing for heteroskedasticity** - Heteroskedasticity basically means that the variance isn't a constant. When heteroskedasticity is present in the cross-sectional data (data that isn't time dependent—this data), it leads to inefficient

estimates and incorrect p-values. Essentially, the results are inaccurate. I could not test for this with my given data set using Excel, so I have no idea if this is playing a factor.

- **Not testing for collinearity** - Collinearity occurs when some of the independent variables are correlated. For example, something like damage rate and return rate would have collinearity because anybody who gets a damaged product is probably going to return it. The presence of collinearity reduces statistical precision of the model.
- **Assuming linearity** - This one is huge. A big problem with this model is that it assumes a linear relationship with the independent and dependent variables. Basically, when the coefficient for review [REDACTED] this means that each review will lead to a GRS [REDACTED]. However, I'm sure that the first review on a product might be worth thousands in GRS, while, for a SKU that already has thousands of reviews, one review would only be worth a few dollars. The one benefit of this model is because we're only considering the top 500, generally all of the products should have a lot of reviews already, but either way this is an important consideration.
- **Excessive number of variables overexplaining the observed data** - Listen, all data has some degree of randomness and "noise" in it. However, if you have 1000 independent variables, it's more likely that a combination of coefficients is going to be able to explain all the randomness in a model. The more variables you have, the better the model will fit the data, but it doesn't necessarily mean that the predictive power of the model will improve. In fact, it might take away from the strength of other variables by explaining some pattern that just isn't there. I don't have an obscene amount of variables in this model, but if I added more I might run into that problem.
- **Sample size** - Working with more data would lead to better results, but I think there would also be diminishing returns with this approach. For one, it would

take way more time, and, in addition to that, the top 500 are already making [REDACTED] the total GRS so the rest of the remaining SKUs [REDACTED]

- **Working with cross-sectional data + averages instead of time-series data -**

This is another huge consideration. My data set doesn't use time as a factor (it's not time-series data). It only uses averages. However, the time of day, or year definitely affects the GRS of a SKU. Even price itself is determinant on time and it changes regularly. The reason I didn't use time-series data is because it is extremely difficult to work with, and would be even more difficult to source from the company.

- **Outliers** - A lot of the top [REDACTED] of [REDACTED]

[REDACTED] thus, these SKUs would wildly influence the regression more than everything else. However, it wouldn't make sense to leave them out of the regression since they are important to consider.

- **Omitted variables** - A couple of things I'd want to control for that would take more time and a lot more data (more than Excel could handle) would be colors, styles, number of images [REDACTED]

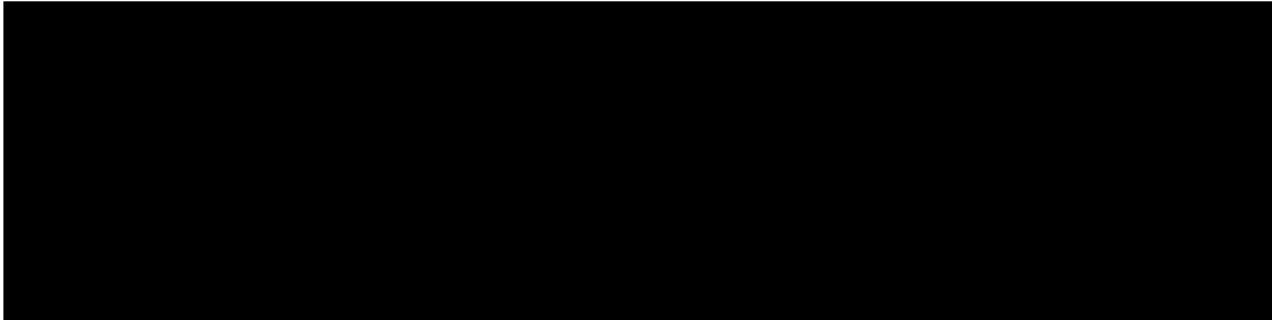
[REDACTED] or the number of [REDACTED] [REDACTED] I do think that these factors could play a significant role, though.

Next Steps

So what should you take away from this study? I think we might have to

[REDACTED] brands BUT, big disclaimer, **within the top** [REDACTED] make assumptions outside of the data set is where you get into trouble, since these results may not replicate for the entire bed class or outside of the top 500. That being said, this is still worth diving deeper into [REDACTED]

I would look into creating a more robust model than mine to see if the results repeat themselves. I would control for the flaws I outlined above, especially by including month over month data and adding in potentially missing variables.



Lastly, I think it would be worthwhile to look at sourcing more wingback beds, since they seem to be a key style that is contributing a lot of revenue on site.

Thank you

- [Meghan Steeves](#) - My manager for connecting me to all the right people and tools I needed to get what I needed for this project. She was my first point of contact on everything
- [Christina Rancan](#) - My mentor who heard the rough draft version of my project before I started working on it and provided my valuable Excel advice on how to turn the attribute data into dummy variables for my model.
- [Melissa Capland](#) - My coworker who completely saved me by helping me maneuver around the scripts to pull the top 500 queen bed SKUs—something that eluded me for days.
- [Tabor Uhlig](#) - A fellow intern who helped me for more than an hour with excel and GBQ to figure out how to pull the data I needed when he too was working on a project.
- [Connor Roope](#) - My coworker who let me use his laptop to run the Excel regressions that I needed.
- [Arjun Prasad](#) - Someone that Meghan connected me to who pointed me in the right direction for my questions.

- [Shuo Zhao](#) - A data scientist that Arjun connected me with that answered all my technical questions about the continuous winner algorithm and pointed me to a slack channel to help me gather the data I needed.
- [Eric Guo](#) - An analyst that Arjun connected me to that answered my high level questions with the continuous winner algorithm and gave me the CW scores for the top 500 queen bed SKUs.
- [Suhas Sarma](#) and [Sarah Lehmann](#) - Two analysts at EB that took time out of their day to explain the counterfactual to me, an important consideration before I started working on this project For [Suhas Sarma](#) specifically, he looked at my data and gave me lots of interesting talking points and insights on my results..
- [Sarah Yip](#) - Someone that replied to my question in the analytics-forum slack channel and gave me a super important script for pulling the data I needed.