Overview

Data Overview

The data provided is a sanitized version of a prior client's data. The customer names and product names do not match up with the invoice data. The data for the Fans category has been generated to fit the case.

Objective

Acme Lighting is a manufacturer of light fixtures. Products have a list price which is generally considered the max price sold to a distributor. Not all distributors pay the distributor list price. Some distributors have negotiated discounts and some products are heavily discounted as they become discontinued. Acme has approached INSIGHT2PROFIT to assist with setting prices. The first step is to get some baseline understanding of the data and then explore one product category, Fans, to create some initial hypotheses.

Questions

1. In discussion w/the client, we heard there is a field called List Price which should capture the list price at the time of the transaction. A product can have multiple list prices at the same point in time. This is intentional and by design. However, we aren't sure which fields(s) create intentional separation of list prices for each ProdId. Help identify those field(s).
    - Investigate ProdId = 528 to determine why the ListPrice varies between ('2012-01-01','2013-01-01')
    - The variation is intentional - the goal is to use a data mining technique to uncover the intentional variation in the data
2. The client is interested in understanding the variation in discounts. What is one of the primary explanatory variables for discounting behavior?
    - Perform a quantitative analysis to identify one of the main variables that explain discounting.
    - UnitPrice is Revenue/Volume
    - Discount is 1-(UnitPrice/ListPrice)
3. We will be meeting with the Product Manager in the Fans category (Category=Fans) and want to have a perspective on how List Prices are set relative to cost and perceived market value. Are there product attributes that could explain a higher or Markup or Margin % from ListPrice?
    - Calculate a measure of list price markup (ListPrice/Cost) or margin % (ListPrice-Cost)/ListPrice and identify variables that may explain how the ListPrice was set or could be set
4. The product manager indicated that List Prices are not consistently managed well within the Fans category. She would like to see a few items that seem to be out of line with expectations.
    - Calculate an expected Margin % at List Price or Markup from List Price for Fans with some additional product attributes to reduce the variation
    - Identify a product that is out of line with expectations
    - Determine what % of the items fall within 10% of the expected price

Approach

1. [x] Load in Raw data
2. [x] Merge data
3. [x] Prepare data for analysis
4. [ ] Load prepared data
5. [ ] Apply the data science technique of **_** to identify what causes a list price to be different for the same product (what fields). Identify the field(s) that explain the intentional differences in pries.
6. [ ] Apply the data science technique of **__** to create top-level segmentation of the business with respect to how prices are discounted. Identify the field captures the majority of the variation in discounting.
7. [ ] Apply the data science technique of **__** to identify product attributes that may have a causal relationship with List Price and Margin % for the Fans Category
8. [ ] Build a basic model using **_** that is used to estimate the effect of those attributes and ultimately estimate a List Price.
    - [ ] Identify a few products that seem over or under-priced compared to the estimated price from the model

Limitations

Indicate any limitations in the analysis in this section

Parameters

Call out any global variables that may be helpful in your analysis to allow an analyst to adjust. This could be a tolerance level

Ideas

What ideas could take the analysis to the next level if more time was allocated to it?

Takeaways

What are the key takeaways from the questions above

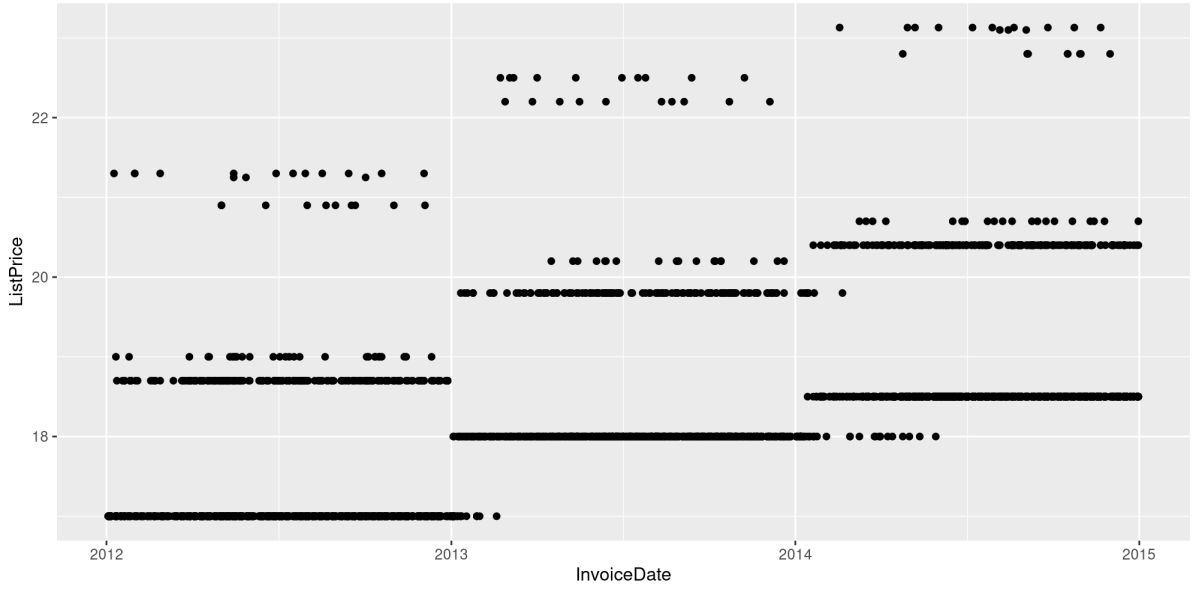
1. Question #1 - ProdId 528 has **_** number of list prices between 2012-01-01 and 2012-12-31 which directly related to the fields of ___
2. Question #2 - One of the largest contributors in the variation in discounting is **__**
3. Question #3 - Fan List Prices appear to be caused by **_ and __**
4. Question #4 - A basic model using ***was created and identified the product*** which is **__** away from it's expected price

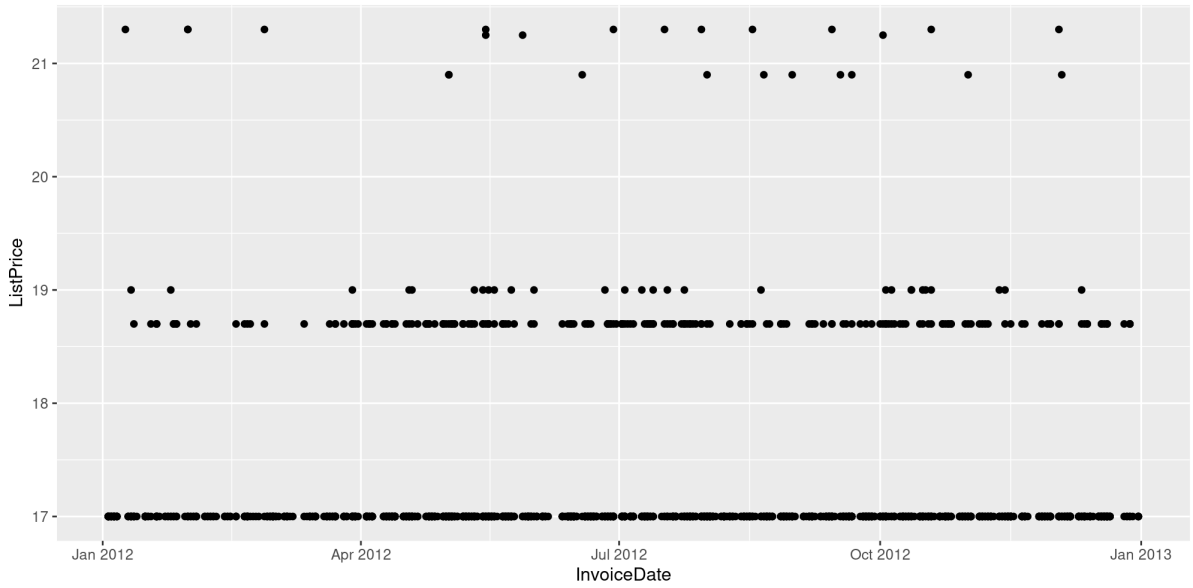
Code

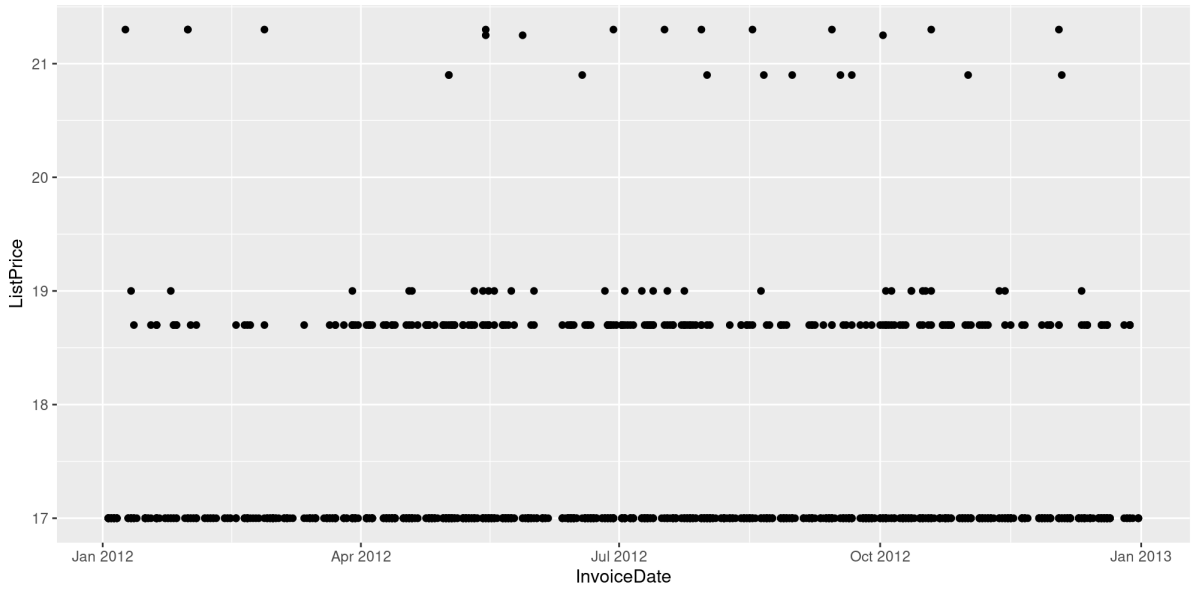
```
▶ target cust_prod_invoices

Warning message in eval(jsub, SDenv, parent.frame()):
“NAs introduced by coercion”
```

Question #1 - what causes this pattern of horizontal stripes of list prices?

Scenario: While exploring the data, the following chart was created. The lines look too perfect to be random variation. What is the cause of the lines? Clearly date, what else?



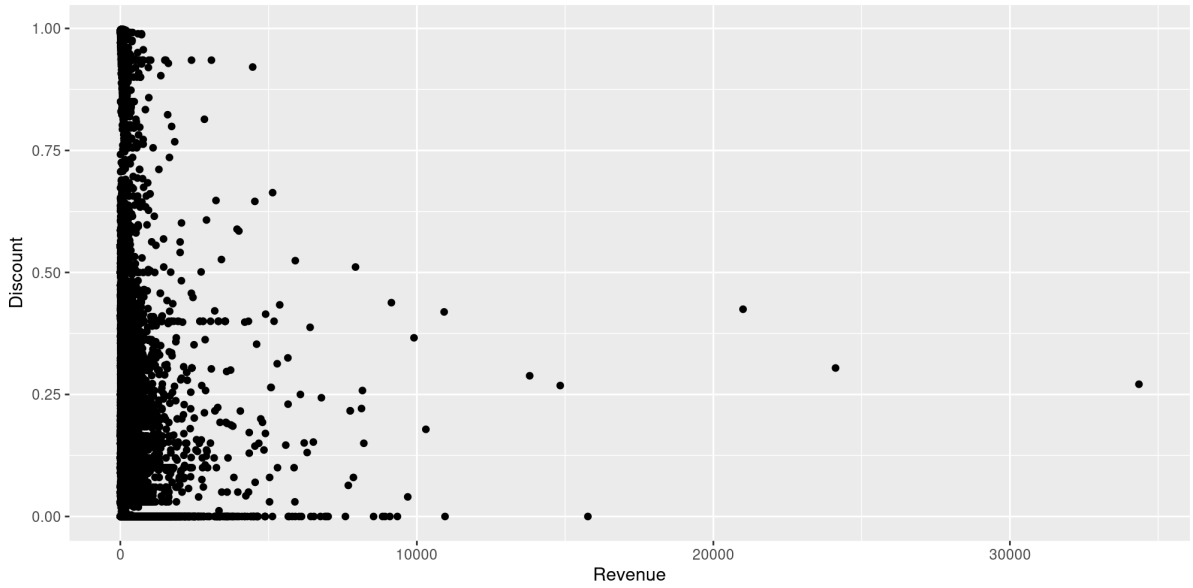


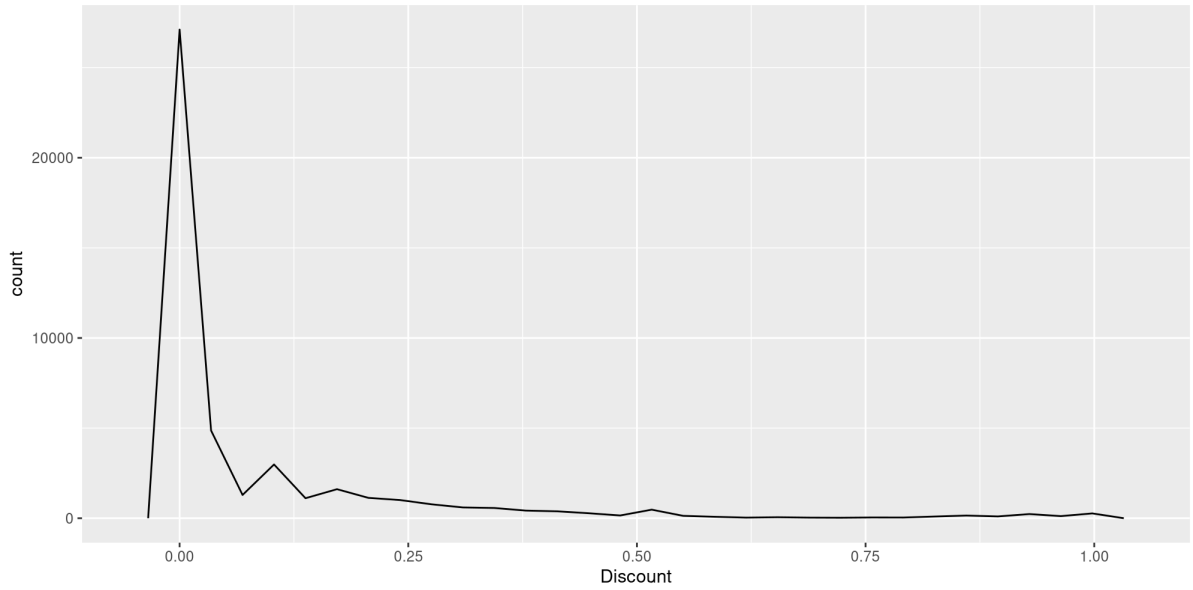


Question/Context: This chart came up in discovery and we were asked to dig in to explain the chart because it looked surprising. What explains the different lines? Show an advanced analytics approach to identify the cause of the different lines. This variation is intentional and by design.

Question #2 - What is one of the main contributors to this discounting behavior?

Scenario: We often look at Discounting to identify market behavior. The chart below looks at one specific month. There are some interesting patterns on the chart which could help with our understanding of the business. Identify some trends/patterns on the chart using advanced analytics.





Question/Context: This is another surprising chart that shows a peak at 0% discount and a few other bumps at 10% and 20%. What data fields help explain this behavior

Question 3 - What attributes affect the List Price of Fans?

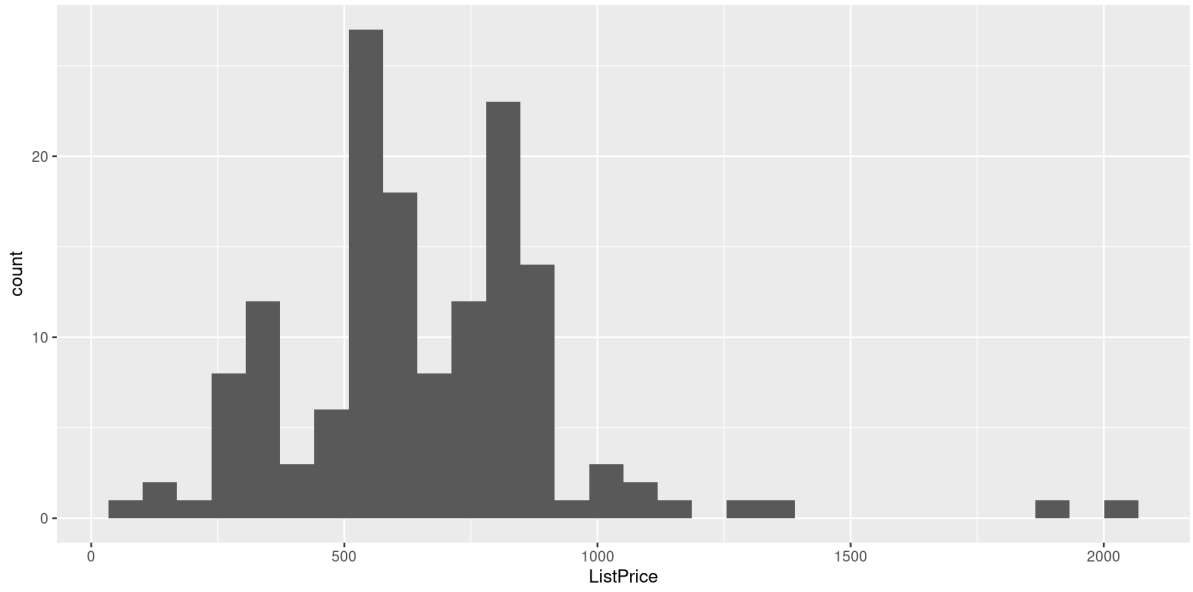
Scenario: We are talking with the Fans Product manager and need to do a deep dive to understand the current state.

A data.table: 5 × 10

ProductNumber	ProductDescription	Line	ProdId	Revenue	MarginPct	ListPrice	Cost	V
<chr>	<chr>	<chr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	
11984	30 Inch Canfield Fan	Builder Fans	11139	182.48	0.5727154	95.90563	40.97900	
11982	30 Inch Canfield Fan	Builder Fans	4241	75.00	0.3294312	109.99945	73.76220	
413SNBU	42 Inch Kichler Basics Premier	NULL	974	46438.04	0.5983581	142.84016	57.37060	
300121SBK	46 Inch Eva Fan	NULL	5059	11696.03	0.7044178	212.57863	62.83446	
S339010NISW	Sterling Manor-Non ES Subway	NULL	16950	8091.19	0.7224642	246.08833	68.29833	

A data.table: 5 × 10

ProductNumber	ProductDescription	Line	ProdId	Revenue	MarginPct	ListPrice	Cost
<chr>	<chr>	<chr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>
15335	70 Inch Monarch Fan	Decorative Fans	375	16230.90	0.9026461	2062.548	200.7971
11989	70 inch Monarch Fan	Decorative Fans	17944	18755.05	0.8949549	1911.532	200.7971
11960	60 Inch Olympia Fan	Decorative Fans	3600	297.00	0.8735967	1361.608	172.1118
12067	60 Inch Ferron Fan	Decorative Fans	12527	110.00	0.8643812	1269.085	172.1118
12058	60 Inch Trevor Fan	Decorative Fans	2605	137.63	0.8521053	1163.746	172.1118



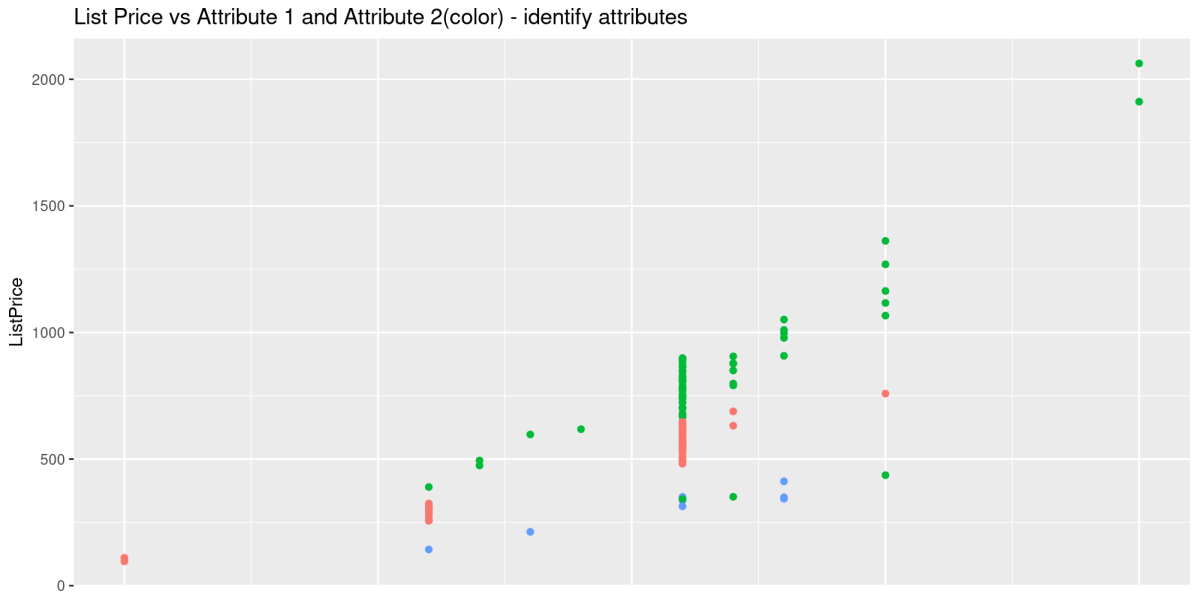
Question/Context: What attributes in the data could have a causal relationship with List Price. Show those variables with a evidence of why.

Question #4 - Can we model a list price using the attributes and identify Fans that need attention?

Scenario: We want to put together a basic model that identifies the difference between actual and expected List Price based on the causal variables.

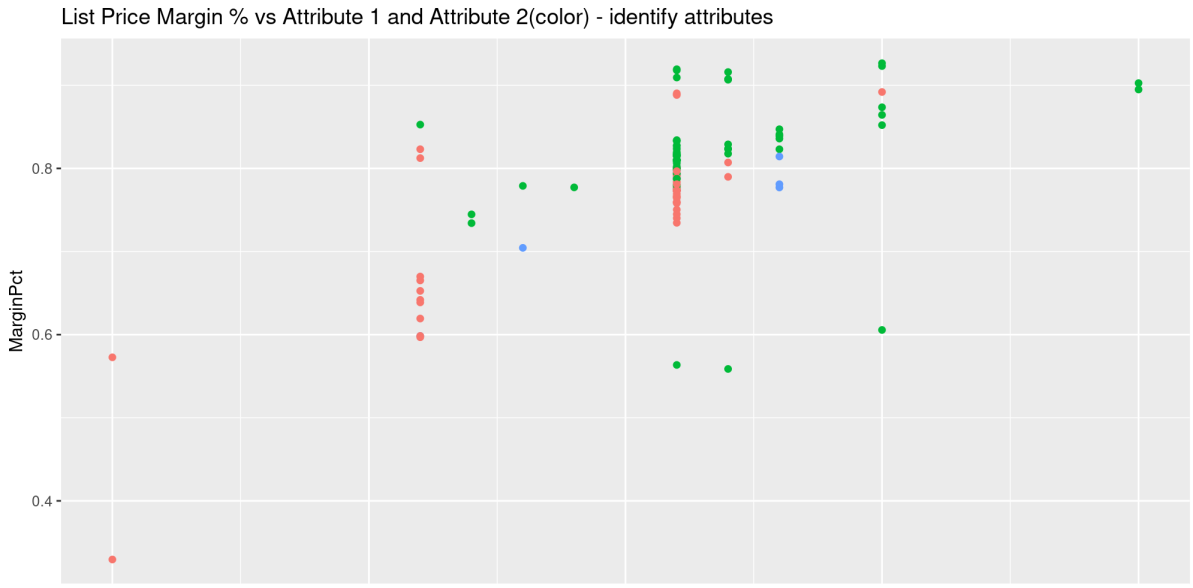
```
Warning message in eval(jsub, SDenv, parent.frame()):
“NAs introduced by coercion”
```

```
Warning message:
“Removed 4 rows containing missing values (geom_point).”
```

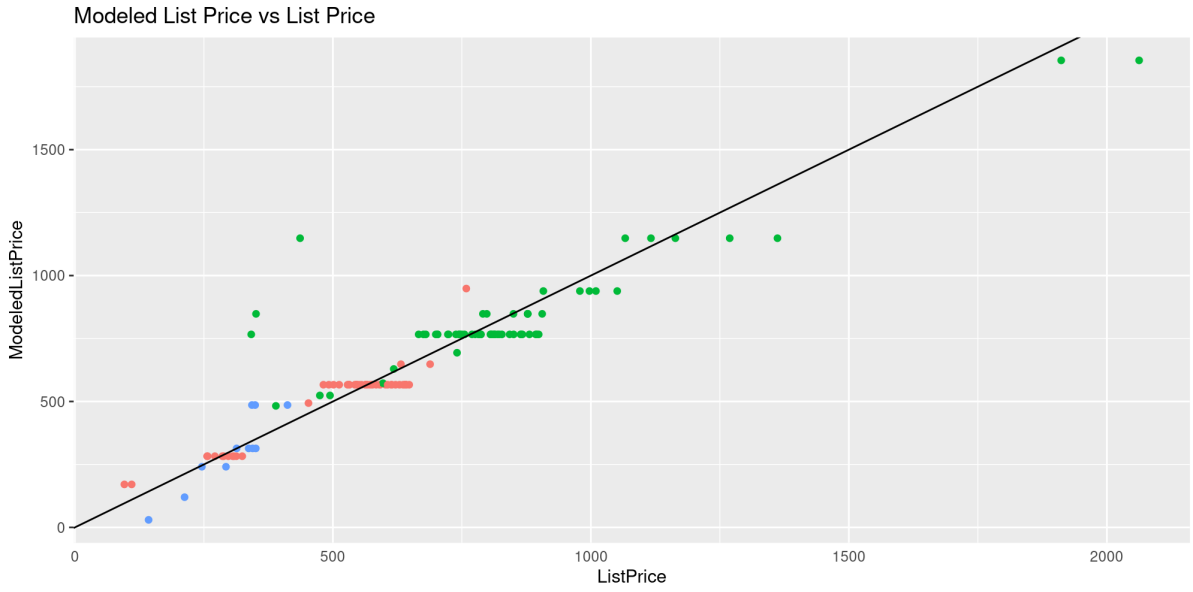


Question/Context: This chart shows what appears to be a relationship between three variables (ListPrice, X, and colour). What are the variables?

```
Warning message:
“Removed 4 rows containing missing values (geom_point).”
```



Question/Context: This chart shows what appears to be a relationship between three variables (Margin %, X, and colour). What are the variables? Are they the same as above?



Question/Context: Create any model that estimates a list price (ModeledListPrice) from the data. Show the model output and a graph

```
[1] "73% of the fans are within 10% of the modeled prices"
```

Question/Context: Show some measure of model 'fit' that shows how similar our modeled price is to the actual list price

A data.table: 5 × 12

ProductNumber	ProductDescription	Line	ProdId	Revenue	MarginPct	ListPrice	Cost
<chr>	<chr>	<chr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>
15366	52 Inch Fitch Fan	Decorative Fans	19255	399.35	0.5635303	341.75005	149.1636
11984	30 Inch Canfield Fan	Builder Fans	11139	182.48	0.5727154	95.90563	40.9790
11982	30 Inch Canfield Fan	Builder Fans	4241	75.00	0.3294312	109.99945	73.7622
11973	54 Inch Bellamy Fan	Decorative Fans	4594	62.00	0.5587096	351.01740	154.9006
11959	60 Inch Olympia Fan	Decorative Fans	16163	5.00	0.6056268	436.41855	172.1118

Question/Context: Show a few fans that appear to be very different from the ModeledListPrice. What are some possible causes? (It can be from any model -- not expecting it to be the same model as above)