Stats 504 Assignment 1: Buying a laptop from eBay

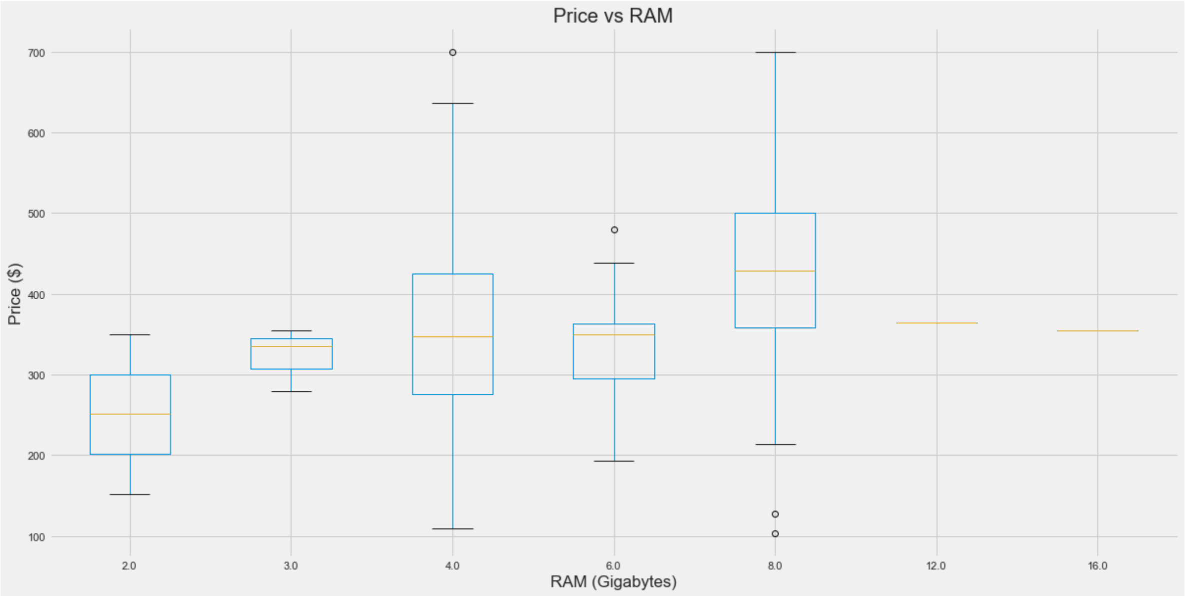
Introduction

This report analyzes laptop prices, in an attempt to understand what features of a laptop influence its selling price. The findings from this analysis are aimed to enable the client to select a favorable laptop deal on EBay. The report also addresses specific questions of interest such as the influence on price by the presence of a Solid State Drive (SSD) and the mode of buying – auction or Buy-It-Now (BIN). The choice of model for this analysis draws from real-world intuition that laptops with better specifications tend to be priced higher, and prices are likely to grow in a linear fashion. Following the analysis, the report also makes laptop recommendations which are considered "good deals" and also describes how BIN and SSD options raise the price of a laptop.

Methods

The goal of the analysis is to help depict and understand the features that influence laptop prices, and provide concrete evidence in identifying a suitable recommendation for a laptop that the client can purchase. Exploratory analysis on the data points towards linear trends in the data, and this is in adherence to real-world intuition. These box plots indicate a positive trend, where the median price of laptops with increasing levels of RAM and GHz, is growing roughly linearly. Owing to this observation, a linear model seemed most appropriate to capture the necessary information from the data.

Figure 1: Seemingly linear increase in price, with a linear increase in RAM size



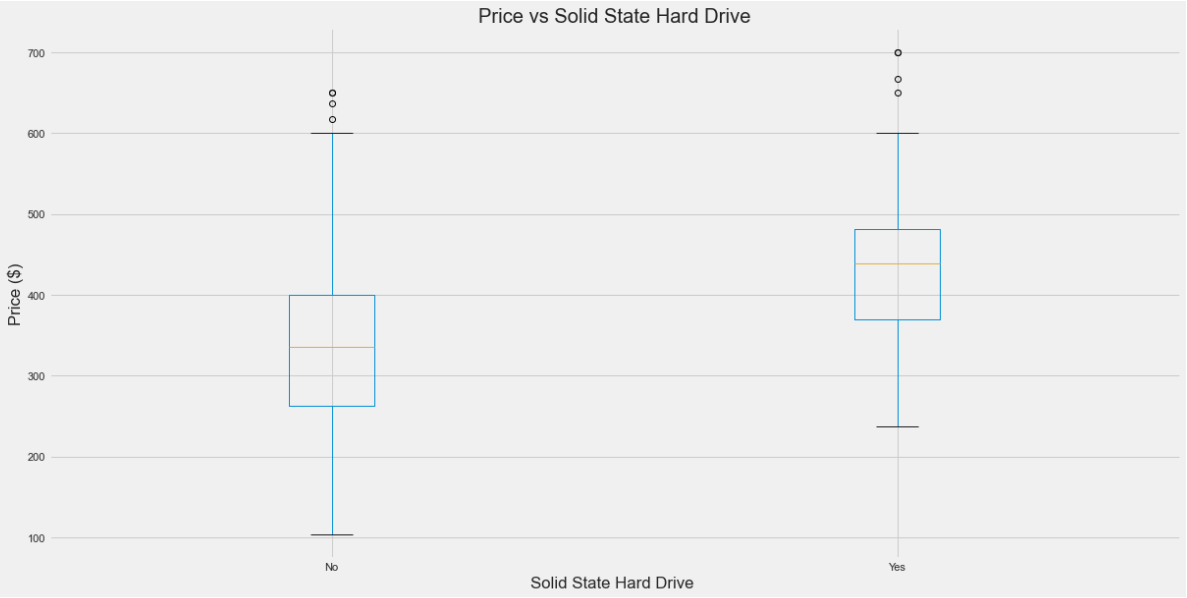


Figure 2: Laptops with Solid State Hard Drives tend to be more expensive than ones without

A Linear Regression Model attempts to fit the best possible straight line through all the data points, such that if given the features of the laptop like RAM, GHz, HD, etc we can (with some level of confidence) estimate the price of a laptop with those specifications, as a point falling on the fitted line. While this model is easily interpretable (discussed in the Results section), it does come with its share of complexities and limitations. If the data were to have inherently non-linear trends, like exponential growth in laptop prices, then this model would not be able to sufficiently capture those trends resulting in a poorly performing model. The model, like most models, also relies heavily on the data that is used to build it. Much care must be taken in vetting the data for the analysis, as even a few outliers can affect the model by a lot, resulting in very different outcomes. All of these were accounted for by performing model diagnostics, and are discussed further in the appendix.

Results

The data provided by the client had details about laptop specifications as well as information about its sale on EBay. Each row of the data represents a different laptop, corresponding to its specific details. However, some rows had missing values (see Table 1 for proportions) or anomalous values (laptop prices less than $5). The data was cleaned using appropriate statistical methods. The anomalous rows were discarded as they were not large in number, and for the missing values, the mode was used. The mode is the most frequently occurring value in the dataset, so this meant missing values were filled in with a value that was most common for that feature (See Appendix for more details). This strategy seemed appropriate for the data at hand due to its discrete nature. Since features like RAM or GHz can only take up certain values and are not truly numerical, using the most commonly occurring value is reasonable. The regression analysis was finally performed on 215 different laptops with each laptop having 7 different features which are further described below.

Feature	**Median (25th, 75th percentile) or Percentage**
Sold (%)	73.02%
Price ($)	361 (300, 450)
Processing Speed (GHz) (%) – 22.2% missing	
2.5	63.72%
2.6	25.11%
2.7	8.88%
2.8	0.93%
3.2	1.39%
RAM (GB) (%) – 19.09% missing	
2	0.93%
4	56.74%
6	7.44%
8	32.55%
12	0.46%
16	0.46%
Hard Disk (GB) **31.36% missing**	300 (128, 320)
Solid State Drive (%)	37.20%
Buy-It-Now (%)	53.02%

Table 1: Baseline table indicating summary statistics of data used in the analysis

Multiple variations of regression models were fitted, but only the (subjective) best one is discussed here. The table below describes the effects of statistically significant laptop features on the price.

Feature	Feature Change	Price Change (95% Confidence Interval)
Gigahertz (GHz)	+ 0.1 GHz	+17.00 (32.23 307.80)
RAM (GB)	+ 1 GB	+9.81 (2.51 17.11)
Solid State Drive (SSD)	No -> Yes	+87.97 (47.50 128.44)
Buy-It-Now (BIN)	No -> Yes	+67.12 (39.35 94.89)

Table 2: Price change that is expected to be seen from upgrading a feature

As the table suggests, SSD, BIN, Gigahertz, and RAM, were the features that most influenced the price of a laptop, while Hard Disk Space was deemed as mostly irrelevant to the price by the model. SSD was the feature that influenced price the most ($87.97). To break it down further, if there were two exactly similar features laptops, one with SSD and one without, the one with the SSD would cost $87.97 more. Other features can be interpreted similarly from Table 2.

This model, while being the most sensible fit for the data, was unfortunately not flexible enough to answer some questions regarding the effect of SSD storage on the Hard Disk (HD) Space. Adding an "interaction term" to this model, as a way of modeling the relationship between SSD and HD, makes the updated model unstable and the ability to perform further accurate inferential analysis is lost. While we cannot guarantee a high level of certainty for the following results, it may still be useful to know them. As per the new model, an SSD laptop with an additional 1 GB of HD space would cost $31.87 less than a laptop with the exact specifications for other features. The new model seems to indicate that SSD storage is inconsequential which is intuitively contradictory to what the boxplots and t-tests suggest, which is also another reason the former model was preferred.

In order to make recommendations of good deals for the client, the model can be used to compute the expected price of every laptop and filter desirable ones from the set of laptops that cost lower than what the model suggests is their "true" price. The recommendations can further be filtered based on requirements like presence of SSD, large HD space, BIN option, etc. The following table presents the best deals for the client to consider.

#	Price ($)	GHz	RAM (GB)	BIN Available	SSD	HD Space (GB)	Expected Saving ($)
23	565.00	2.5	8	Yes	Yes	240	90.63
30	500.00	2.7	8	No	Yes	160	58.35
53	564.95	2.5	8	Yes	Yes	128	90.03
100	579.00	2.7	8	No	Yes	160	137.35

Table 3: Recommendations for unsold laptops that are good bargains

All of these recommendations were chosen such that their price on EBay was cheaper as compared to the expected value predicted by the fitted model. This difference in price in depicted in the last column as a measure of how “good” of a deal it is. Since the client preferred to have an SSD, all recommendations have that feature available. The client was undecided about the BIN option, but preferred to not partake in an auction only on the condition that it did not affect the price. The analysis revealed that using an auction may result in landing a better deal, which is why the two recommendations (#30 and #100) have been made. All recommendations have large HD space available as that was another requirement. Considering all of the client’s requirements, these recommendations met most of the checkboxes, as well as had a high expected saving, and are expected to be ideal buys for the client.

Conclusion

This report presents the results of an analysis performed by on laptop prices, with the objective of trying to understand what features influence price. It also addresses the key questions posed by the client regarding specific features like SSD, and BIN and their influence on HD space. The presence of SSD and BIN options increase a laptop price by $87.97 and $67.12 respectively, and the cost of adding HD space is statistically insignificant with respect to whether or not the memory type is SSD. While these results are expected to be useful, it must be noted that the model comes with a fair share of limitations. The strategy for filling in missing data could have strongly biased or affected the model, so re-running the analysis with a better dataset may yield different (and perhaps better) results. Finally, the laptop recommendations made in the results section aim to strike a balance between subjectively trying to address the client’s needs, while at the same time use the information provided by the model. These recommendations could change based on the reader’s interpretation, and the reader is encouraged to look at the appendix for a larger candidate set of potential laptop recommendations.