Nov. 27, 2024

ISL FA24 Project - Task #1 Technical Report

1. **Data description: how many observations? How many features? What type of features?**

   The data originally had 1,275 observations with 22 features consisting of 9 numerical and 13 categorical features.

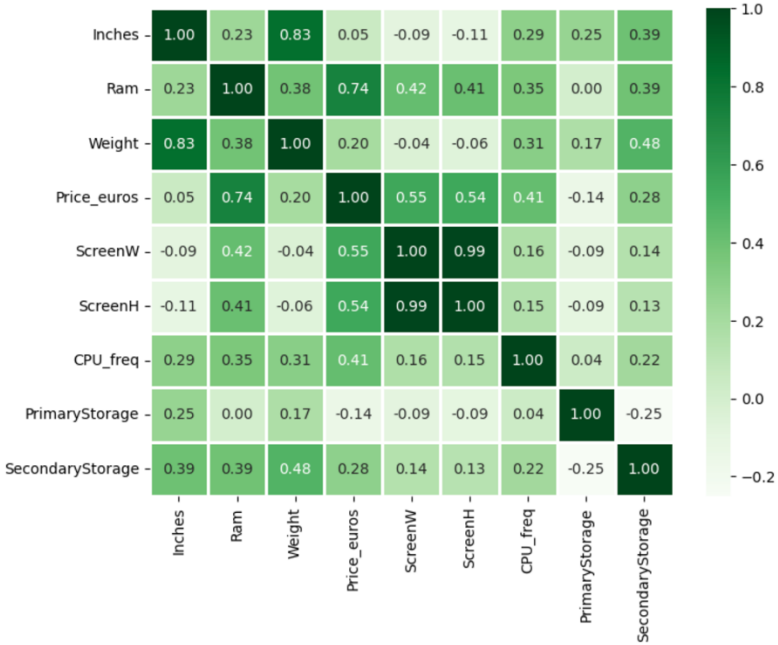
2. **Data preprocessing: are there any null values or outliers? How did you deal with them? How did you handle scaling?**

   There are 7 null values: 4 from Price Euros, 2 from Retina Display, 1 from CPU frequency. There is 1 duplicate row. Any row with at least one null value and rows that had duplicates were deleted since it was not a large portion of data being deleted. There was a category found in the Touchscreen variable which had its row removed as well.

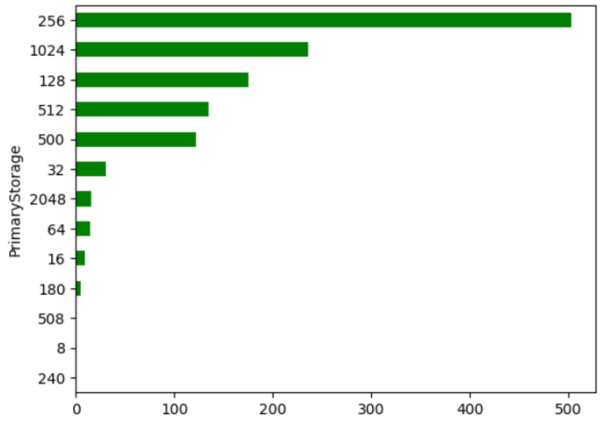
   The outliers in this data were harder to notice since the quantitative variables naturally had a large standard deviation. To find outliers, boxplots & description tables were used to view the data's distribution for each quantitative variable. The criteria for these outliers were laptop: weights > 5kg, screen heights > 3000 pixels, & ram < 3 GB. These criteria removed outliers for values that do not occur naturally or had minimal observations. After cleaning, the dataset consisted of 1248 observations and 22 features.

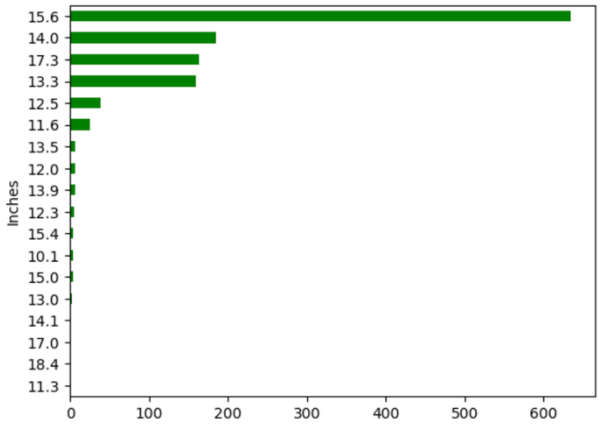
   The dataset was then split into two types where either the quantitative features were scaled or not. However, both contained unscaled dummy variables for the qualitative features. This resulted in resulted in a 1248 x 271 dataset. The dataset used for Ridge, Lasso, & PCA was scaled. However, both datasets were for multi-linear regression.

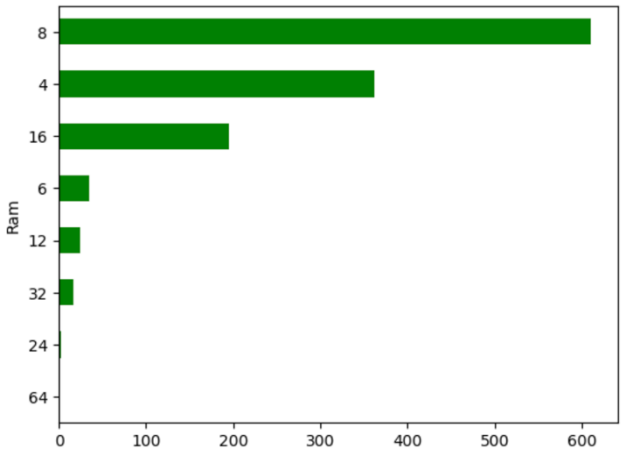
3. Exploratory data analysis: visualize the data with graphs and describe your findings. Did you find any patterns.

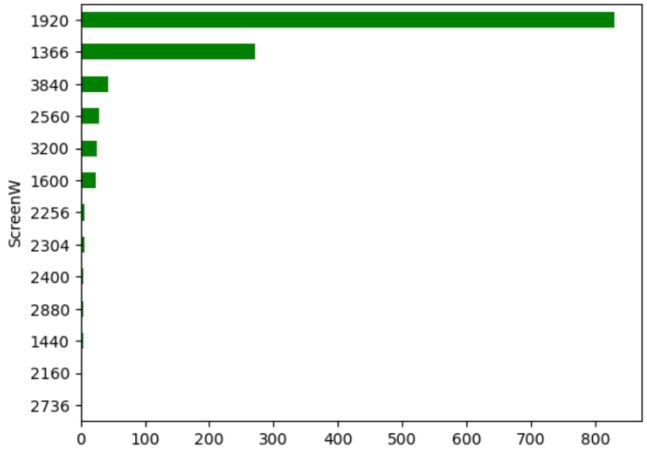


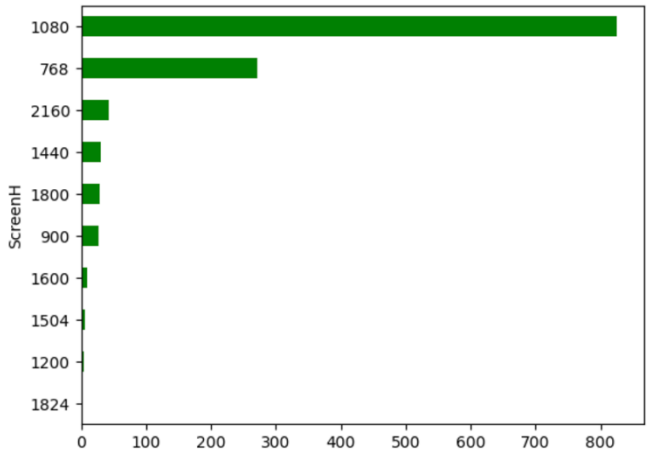
From the correlation matrix Ram, Screen Width, and Screen Height have the most correlation with price. 8 GB of Ram has the most popular RAM size with around 600 samples. Screen Width and Height are highly correlated and have about 800 observations of laptops with 1920x1080 pixels. Inches & Primary Storage size, however, had low correlation with price. The most popular screen size in inches was 15.6 inches and having 256GB of primary storage was favored with 1TB as the next favored storage size

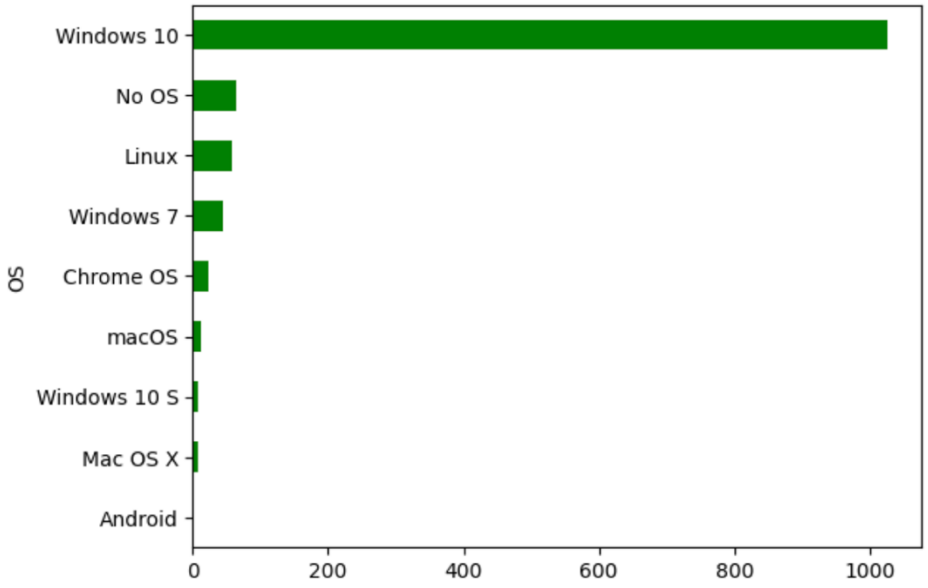




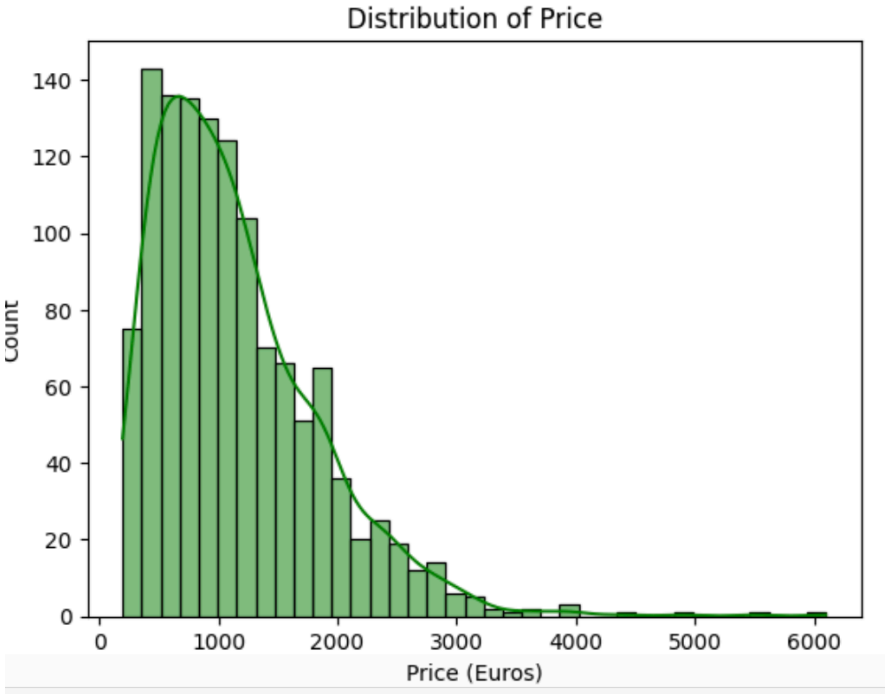


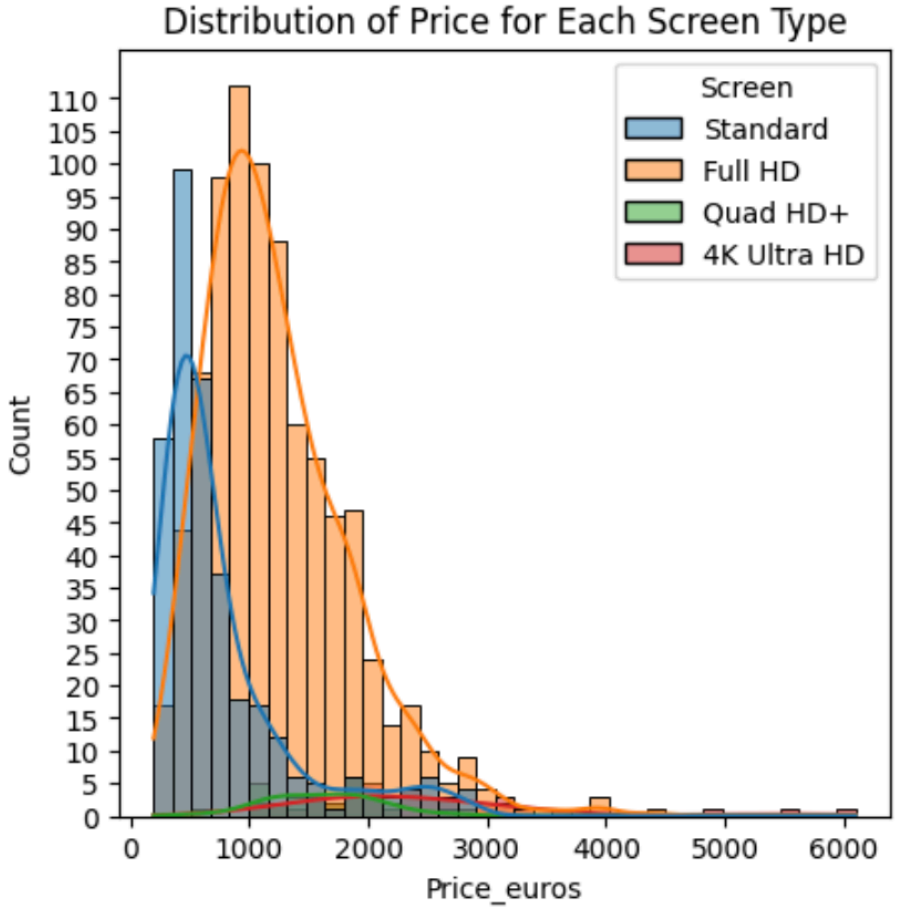






Most of the operating systems are in Windows 10 with 1000 observations. The cost of a laptop averaged around $1000 dollars, but the average prices are different based on the types of screens each laptop has. For a standard laptop resolution, the average price is around $500 dollars. While for a full resolution, the average price is just above $1000 dollars. For Quad and 4K resolutions both have average prices that stay near $2000 dollars.





4. **Model development: state the hyperparameters selected for the models and how/why you selected those hyperparameters**

Linear Regression

- I split my dataset into **two types**: one with scaled numerical features & one without. Both have unscaled dummy variables. When conducting Linear Regression, the unscaled dataset performed better than the scaled.

Ridge Regression

- When performing ridge regression, the scaled dataset was used. There was an improved performance of the model. A list of 100 random alpha values was created to for the alphas parameter in RidgeCV.

- To find the best alpha value, I used the **RidgeCV** class with parameters {alphas , cv=8}. I choose 8 folds since it's recommended to choose between 5-10-folds and 8 folds would be ideal for learning the data without capturing too much variance and increasing computational time.

- For the **Ridge** class, the parameters used were {alpha , fit intercept , tolerance , max iteration , solver}.
  - The alpha parameter used was the value found in RidgeCV, and fit intercept is kept True to include intercept in the model.

  - Tolerance and Max Iteration was tested with these values respectively.
    - tol=[0.01, 0.001, 0.0001] & iter = [100, 1000, 10000]
    - The best values for those parameters were Tolerance = 0.0001 and Max Iteration = 1000

  - Solver was set to 'sag' since this hyperparameter is useful for large observation and feature datasets. Additionally, since we scaled our data, convergence is guaranteed.

Lasso Regression

- The scaled dataset was also used for Lasso. The performance was also similar too Ridge.

- The same set of 100 random alpha values used in Ridge was also used for **LassoCV** to find the best alpha value. The parameters used were { alphas , cv=8 , fit intercept , max iteration , positive }.
  - CV=8 folds was used to test best alpha without capturing too much variance.
  - Fit intercept was kept true to include an intercept during training
  - Max iteration was 100,000 to ensure a convergence of alpha values
  - Positive is kept true to select positive values because we assume some relationship between the features.

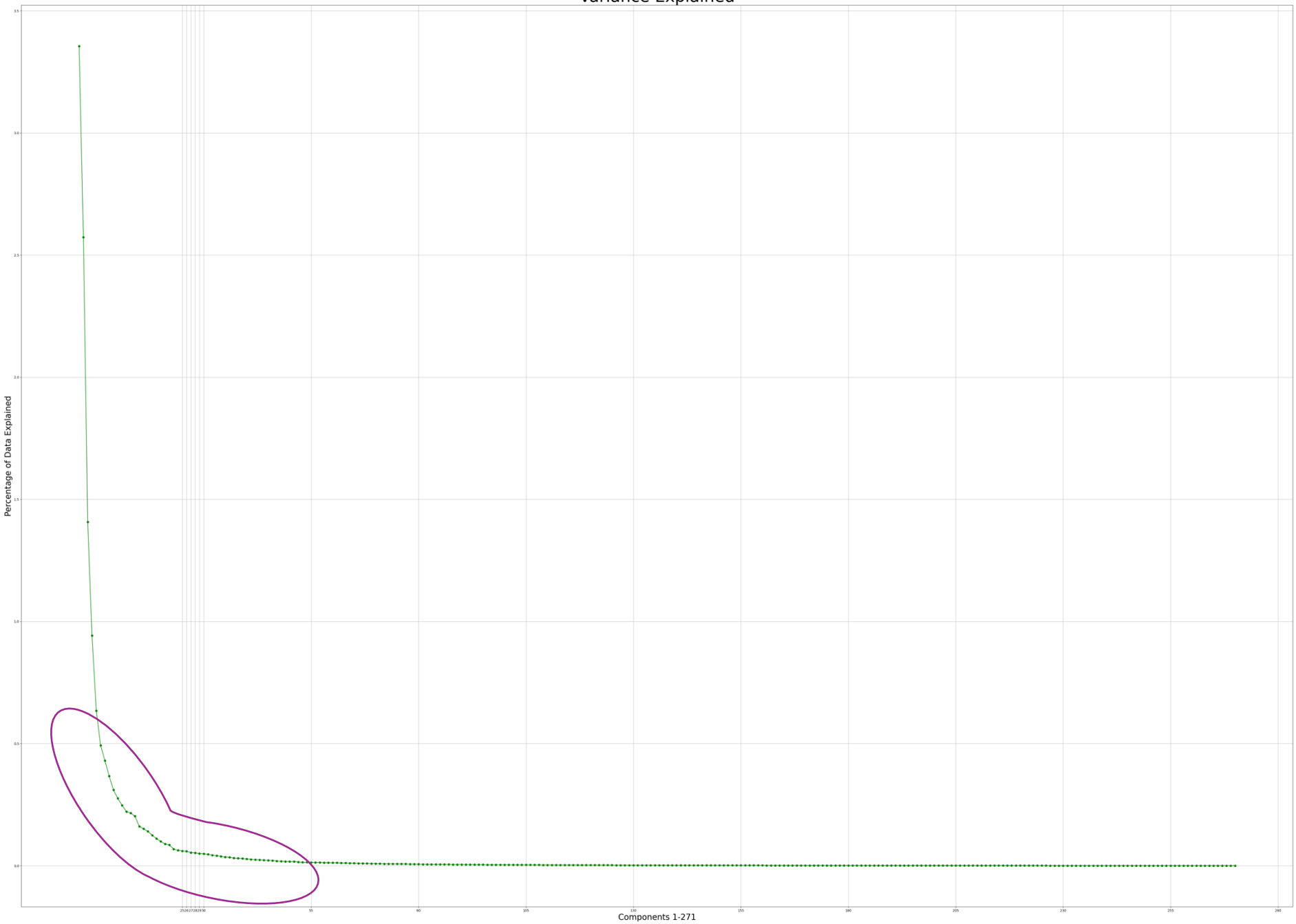
- The parameters for **Lasso** were { alpha , fit intercept , tolerance , max iteration , selection , positive }.
  - The alpha parameter used was the best value found in LassoCV
  - Fit intercept & Positive was kept True.

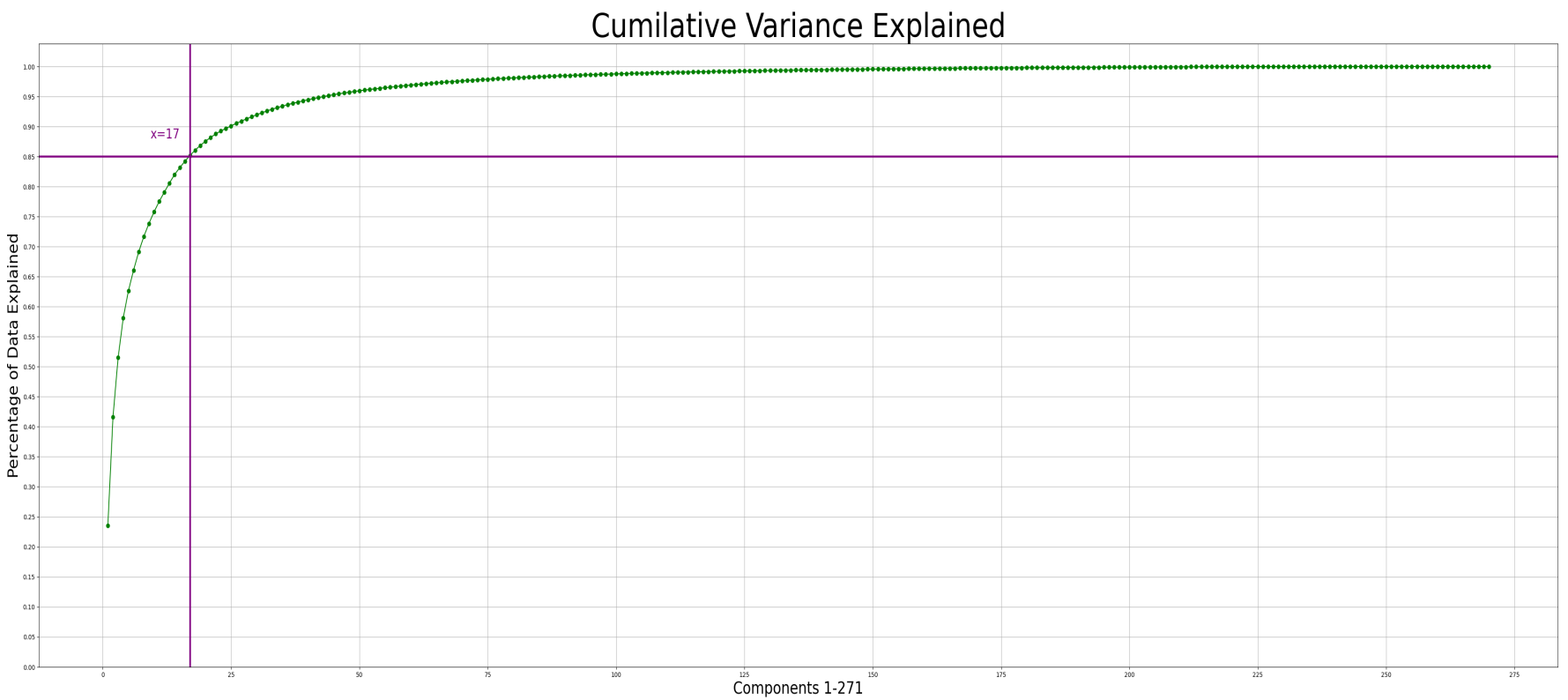
  - Tolerance and Max Iteration was tested with these values respectively.
    - tol=[0.01, 0.001, 0.0001] & iter = [100, 1000, 10000]
    - The best values for those parameters were also Tolerance = 0.0001 and Max Iteration = 1000

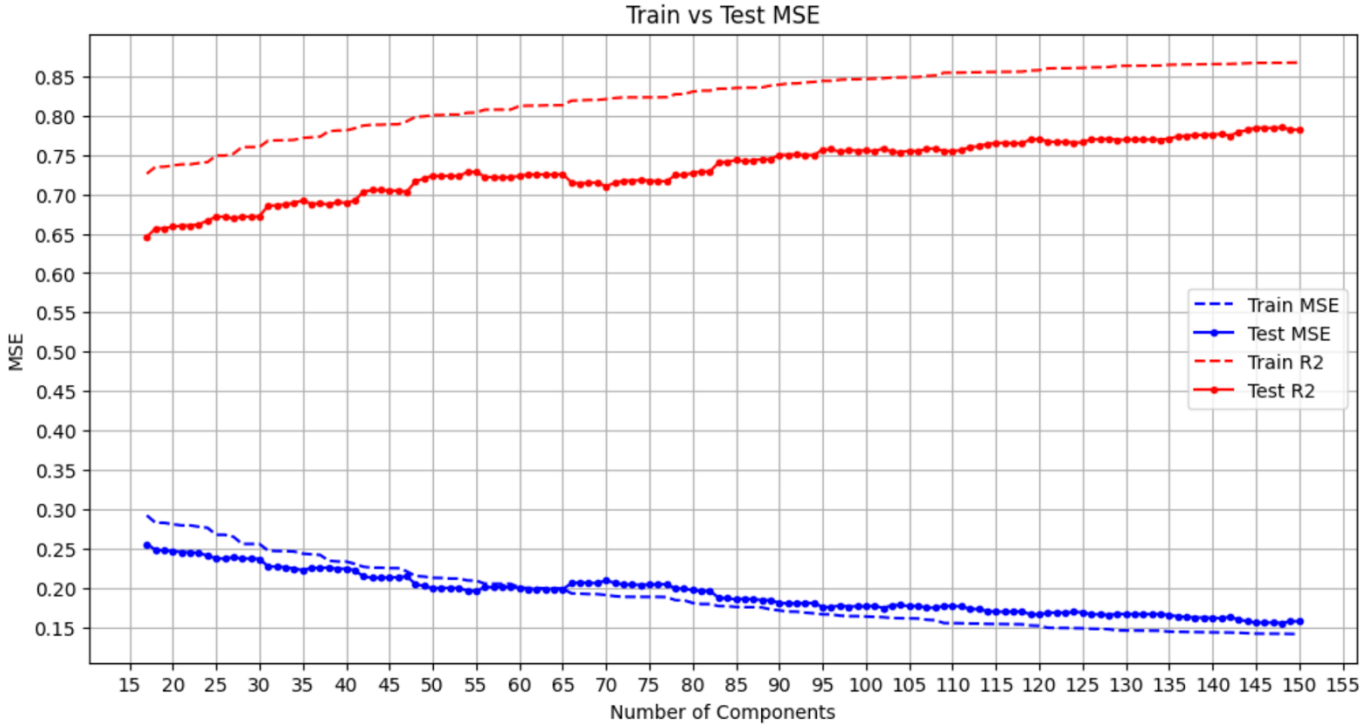
  - Selection was kept as 'cyclic' instead of 'random' so every feature gets iterated through in Lasso.

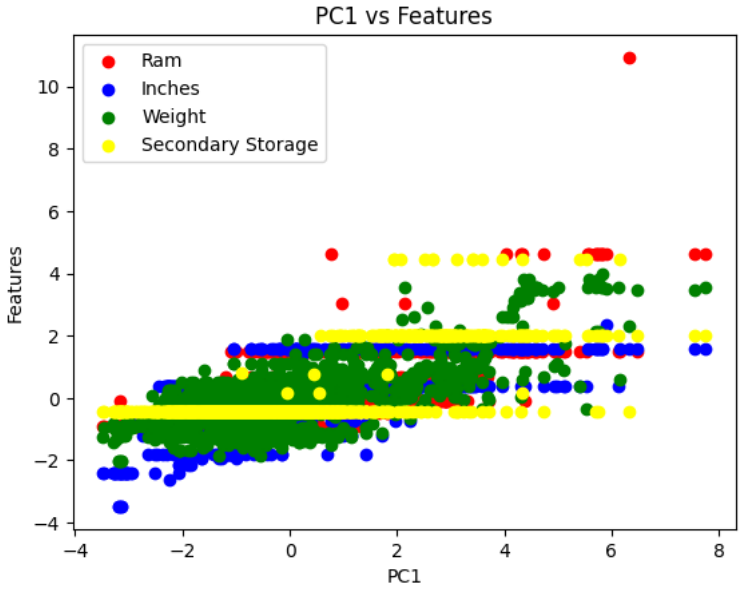
Principal Component Analysis – PCA

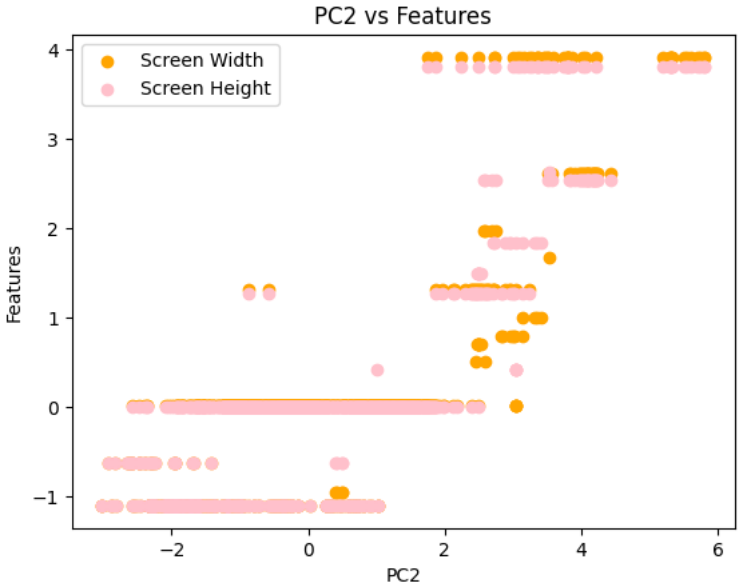
- The only parameter used was n_components = 271, which was the number of columns in x.
- After calculating the eigen vectors , eigen values , and cumulative variance explained – the plots below were used to detect an ideal number of components for a reduced feature space.
- I choose to use the 117 components since I got similar results as ridge and lasso.

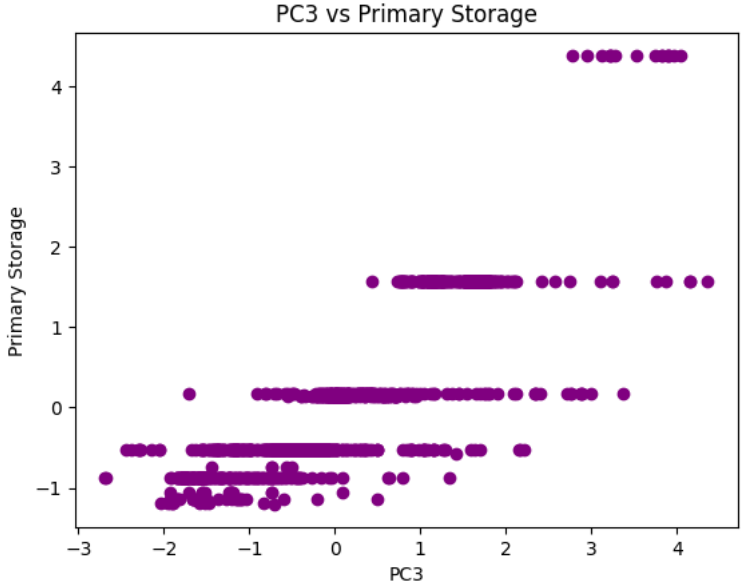

Variance Explained
Percentage of Data Explained
Components 1-271

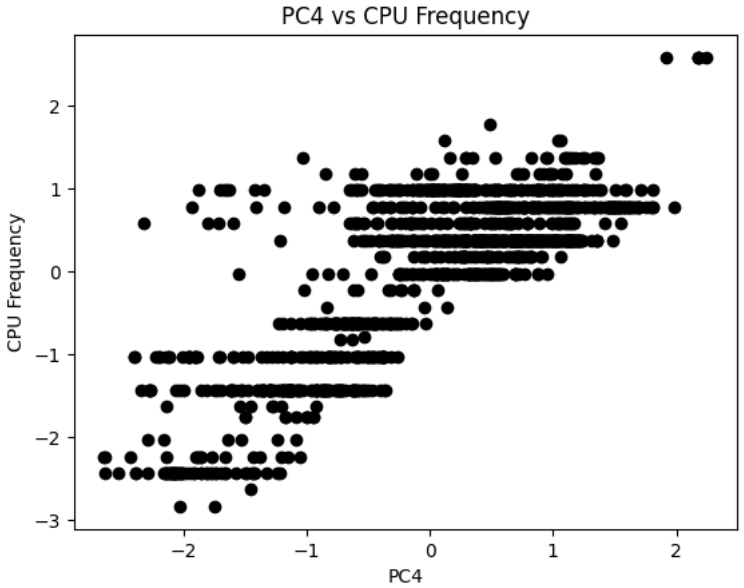

Cumilative Variance Explained
x=17
Percentage of Data Explained
Components 1-271


Train vs Test MSE
MSE
Number of Components
Train MSE
Test MSE
Train R2
Test R2


PC1 vs Features
Ram
Inches
Weight
Secondary Storage
Features
PC1


PC2 vs Features
Screen Width
Screen Height
Features
PC2


PC3 vs Primary Storage
Primary Storage
PC3


PC4 vs CPU Frequency
CPU Frequency
PC4

5. **Performance evaluation: state the model results - accuracy, loss, precision, recall, f1-score, confusion matrix, etc.**

Linear Regression (Unscaled)			
Train MSE	53192.93	**Test MSE**	135668.75
Train R2	0.8974	**Test R2**	0.6139

Linear Regression (Scaled)			
Train MSE	0.1093	**Test MSE**	1.10 e+20
Train R2	0.8974	**Test R2**	-1.52 e+20

Ridge	
Test MSE	0.1563
Test R2	0.7835
Best Alpha Value	1.41747

Lasso	
Test MSE	0.1602
Test R2	0.7780
Best Alpha Value	0.00046

Linear Regression (PCA Data – 117 Components)	
Test MSE	0.1698
Test R2	0.7648

6. **Interpretation: state why a model is performing better/worse. What are the most significant features? Why is the model classifying or clustering to a specific cluster? How does the model results relate to the dataset and the problem?**
   - For the first trial of linear regression the model performed the worst out of all methods for both scaled and unscaled data. Both training R^2 scores produced good results and the MSE for training on scaled data was very good. However, on the test metrics for scaled and unscaled data, the model predicted poorly. Moreover, it predicted worse on the scaled data.
     - The most significant features in both OLS tables were RAM and the four types of screens - 4K HD , Full HD , Quad HD , Standard which agreed with our findings while exploring the dataset.

   - For Ridge, Lasso, & PCA, all methods at feature reduction performed similarly according to their metrics. The alpha values also were low which helped satisfy the dimension reduction properties.

   - However, in PCA the 117 components, which was less than half of the original feature space, was necessary in producing similar results. I choose 117 components because of the metrics it produced and the significance of most of the features in the OLS table. However, I risked overfitting the data by choosing that many components.
     - The most significant features in the PCA OLS table were PC1 to PC8 with PC1 to 4 capturing quantitative variables the best. These can be seen above.
     - PC1 being correlated with 4 quantitative features helped understand the dataset with an eigen value of 3.35.

   - The model performs better after ridge, lasso, and PCA. However, I didn't find any significant improvement of performance between the three. Since my goal with the model was to reduce variance and produce better metrics, Ridge & Lasso could have performed similarly. If I kept only 17 components then it would most likely underfit the model since it had an R2 score around 0.65. Additionally, each component explained very little and required so many components to explain or reduce variance of the data.

   - The metrics from the shrinkage and reduction models indicate there is low multicollinearity between the features. However, since these methods improved our metrics of linear regression not all features are independent from each other but also don't require such strong reduction techniques. The dataset may be too complex for linear regression and may perform better under other models such as Regression Decision Trees or SVM.

7. **Conclusions: summarize the project.**
   - After data cleaning and preprocessing, the data consisted of 271 features. The linear regression model was found to perform poorly with both scaled and unscaled datasets. In the OLS table the most significant features were Ram and the 4 different screen resolution types. After Ridge, Lasso, the linear model's performance improved consisting with the same significant features. The feature space was also reduced using PCA and found that 117 components was needed to achieve similar results as in Ridge and Lasso Regressions and the first 8 components were highly significant. However, only 17 components were necessary to explain 85% of the variance. With all three shrinking/reduction methods, the metrics of our model stayed consistent with an MSE $\approx$ 0.16 and R2 $\approx$ 0.76. This revealed that the feature space may not be highly correlated, or that the data is too complex for a linear model. In future projects, the model may perform better with decision trees or neural networks.

8. References

James, G. *et al.* (2023) *An Introduction to Statistical Learning with Applications in Python*.

Sklearn Developers (no date a) *Lasso*, *scikit*. Available at: https://scikit-learn.org/dev/modules/generated/sklearn.linear_model.Lasso.html (Accessed: 01 December 2024).

Sklearn Developers (no date b) *LASSOCV*, *scikit*. Available at: https://scikit-learn.org/dev/modules/generated/sklearn.linear_model.LassoCV.html (Accessed: 01 December 2024).

Sklearn Developers (no date c) *Ridge*, *scikit*. Available at: https://scikit-learn.org/dev/modules/generated/sklearn.linear_model.Ridge.html (Accessed: 01 December 2024).

Sklearn Developers (no date d) *RIDGECV*, *scikit*. Available at: https://scikit-learn.org/dev/modules/generated/sklearn.linear_model.RidgeCV.html (Accessed: 01 December 2024).