CIS 3252 Business Intelligence, Section 1, Spring 2021

May 22, 2021

Professor Fadi Batarseh

Jacob Yim

Table of Contents

Introduction..... *2*

Data Analysis..... *4*

Results..... *6*

Conclusion..... *11*

Reflection..... *12*

Introduction

As an individual who is interested in details of property management, I wanted to explore the knowledge that revolved around real estate. One could possibly have large profits by investing some of their disposable income onto property investments. Since I had no prior specific knowledge in incoming data, I would be doing a lot of exploration before making any kind of analysis. While looking at the data column description, I had my eyes set on learning impact of the sales price in real estate.

When thinking about the individuals that are involved with property, I can deduct the professionals that are the stakeholders in my domain. If one were to analyze on the size of a property and it's given price, I know that the sectors that would be interested in this work would include landowners, neighboring residents, environmental specialists, building designers and consultants, bank lenders, investors, and governmental utility agencies. From the business perspective, I can identify the three main professionals, which are realtor, property managers, and brokers.

The zip file included from the Kaggle has one description text file, test data, and train data. The original purpose was intended for individuals to practice their knowledge of machine learning techniques alongside data cleaning and exploration. I will be utilizing the train dataset to examine the variables and contents included. The dataset has 81 different columns that describe the variations of houses within the Ames, Iowa location and row size of 1460 samples. The columns have both categorical and quantitative data types ranging of all four different types of measurements (nominal, ordinal, interval, ratio). The big target variable of interest in this dataset is the "salesprice," for it is the main dependent variable for my causations. As I was looking through this data, I had come in with the knowledge that the values would be messy; however,

the data was utilized for a competition for regression models. Although there are null values, it will not impact the weight of this analysis.

The most important prospect of any business is the sales, and in property management, the importance of the price caries a similar value. This analysis will investigate the sales price of the homes and consider its impact with the other descriptive variables in the dataset. Since there are many variables that are required to consider the price for a home, this analysis will show us the key aspects that individuals consider when either purchasing or listing a house. My hypothesis is that the sales price is highly impacted by the neighborhood location, house size, house condition, and house type.

Data Analysis

The in-depth analysis is a two-part investigation that first starts out with comparing the dataframe variables before exploratory data analysis. While looking at the shape of both datasets, I noticed that the train dataset had one extra column than the test dataset. Since there are at least 80 columns, it could be possible that both of the raw data could have different variations or orders of the columns. To compare both datasets, I concatenated them into one dataframe to check if they were likeable. At the same time, this allowed me to look at their data types in one central location if needed. Since there were no different variations or orders of the columns after merging, I compared the two datasets to see which column in the train dataset was unique (Figure 1). Although there are 81 columns in the train dataframe, I checked with the 80 columns from the test dataframe. With no resulting output, I knew that the last column in the train dataframe is the unique column within the two datasets; this columns is named "salesprice."

Figure 1 Dataset Comparison

```
#This is how we know the last column in df_train is unique
for x in range(len(df_test.columns)):
  if df_train.columns[x] == df_test.columns[x]:
    continue
  else:
    print(df_train.columns[x]) #print the column if there is a unique column
```

The next part to make sure my data is clean is to evaluate the duplicate rows. Sometime the listener may make a mistake and accidentally record duplicate values. By using the pandas duplicated method, I was able to confirm that there were no duplicate rows in my dataset. Now

that my dataset is clean of duplicate values, I can now evaluate the sales price by using NumPy describe method.

Figure 2 SalePrice Describe

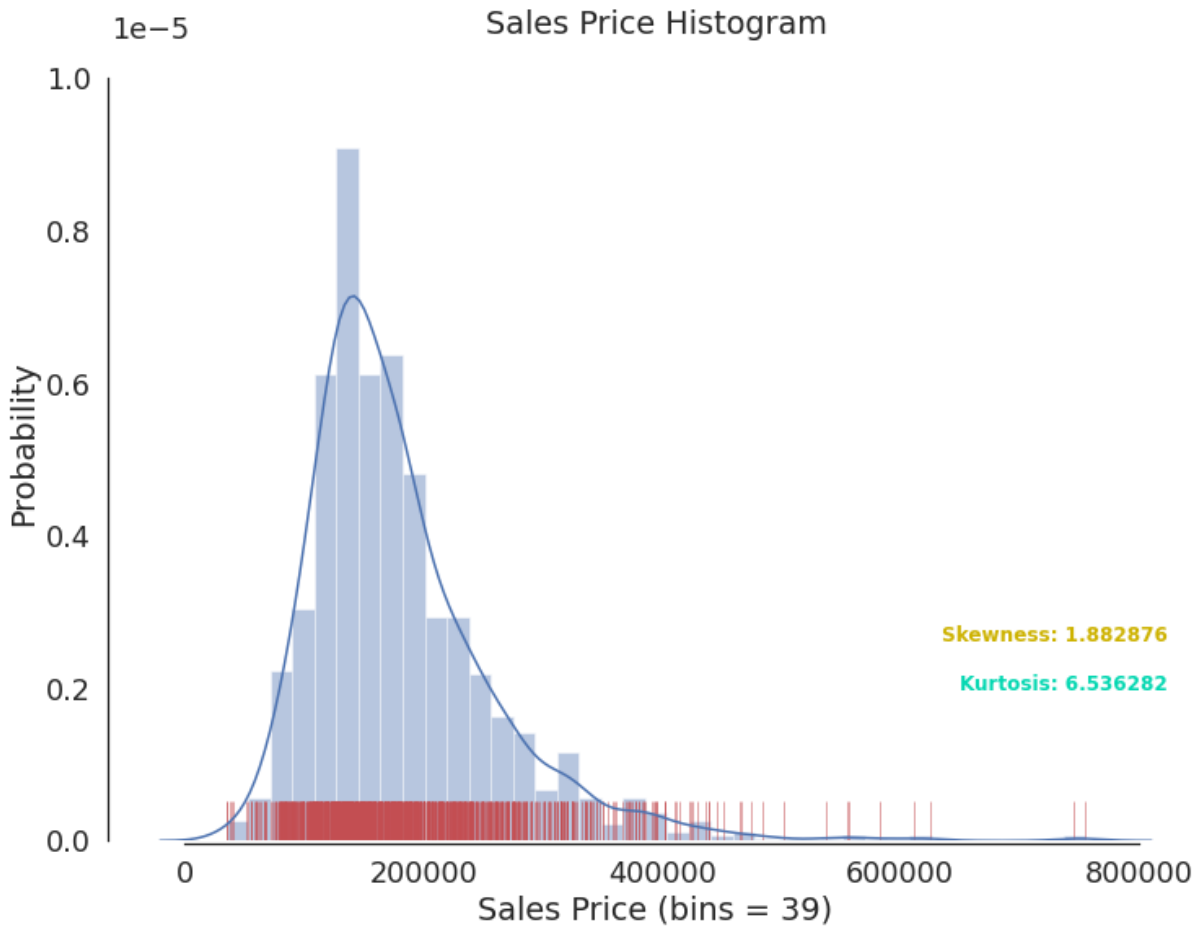
```
[ ]  #describe the sales price
     df_train['SalePrice'].describe()

     count      1460.000000
     mean     180921.195890
     std       79442.502883
     min       34900.000000
     25%      129975.000000
     50%      163000.000000
     75%      214000.000000
     max      755000.000000
     Name: SalePrice, dtype: float64
```

Our first observation in the describe method (Figure 2) is the minimum positive value. It is important to note that the data of the sales price do not contain any negative values, and this is good because it is impossible to have negative profit; this has the same reason as to why one cannot profit a negative sum of money. my mean seems to lie roughly around the 50% and 75% quartile of my dataset which implies a possible skew in the normal distribution.

With a good basis of the sales price, I was ready to move into the EDA portion of the analysis. The describe method would lead into a frequency distribution to visualize the level of skew and kurtosis, and further compare the impact of the sales price amongst the other descriptive columns in the dataset. Using plots that involve correlations would help us to understand relationships, and box plots would help to imagine the central tendency of multi variates. Using these graphs will help to solidify my understanding of the sales price relationships.

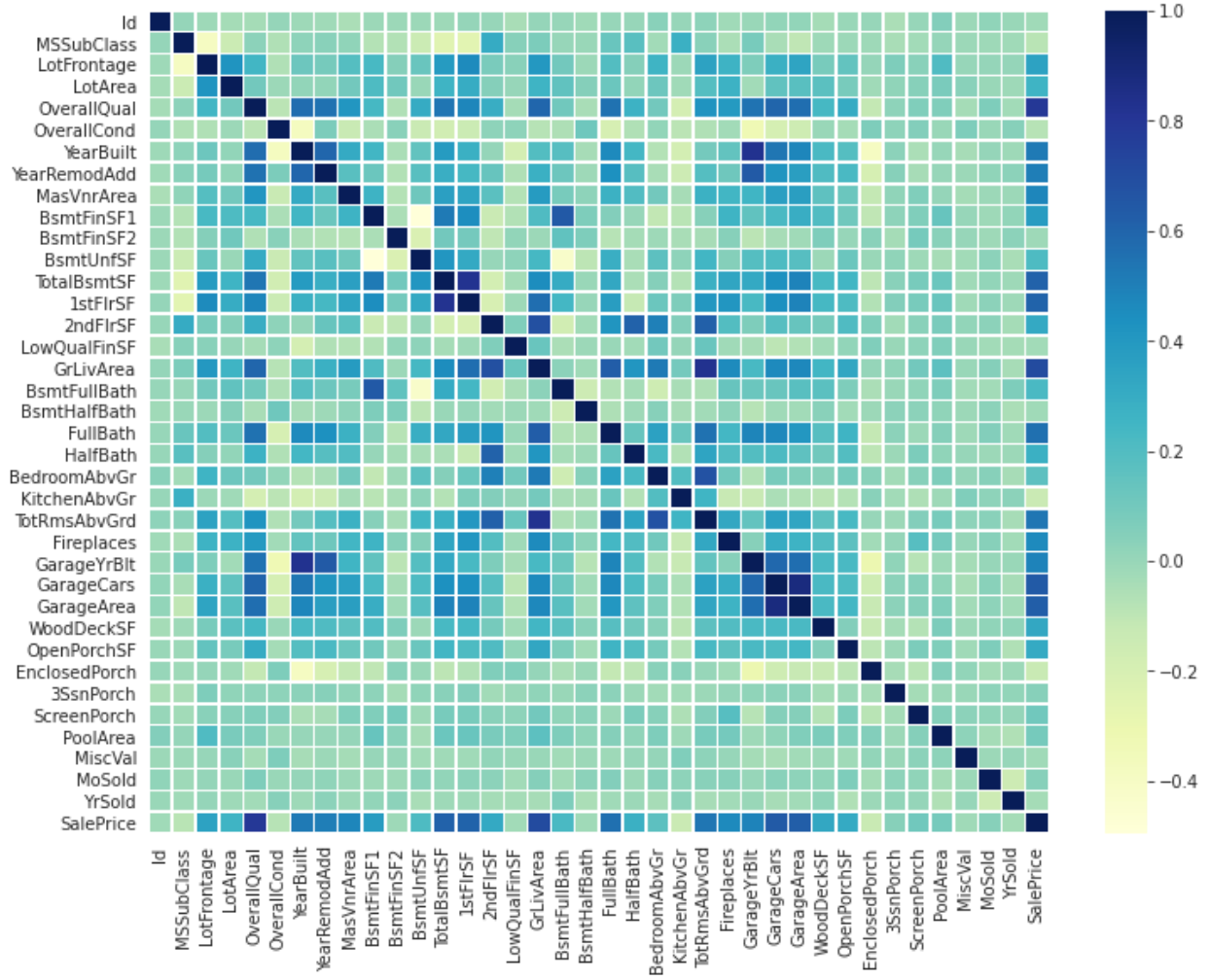
Results

Figure 3 Histogram / Rug plot



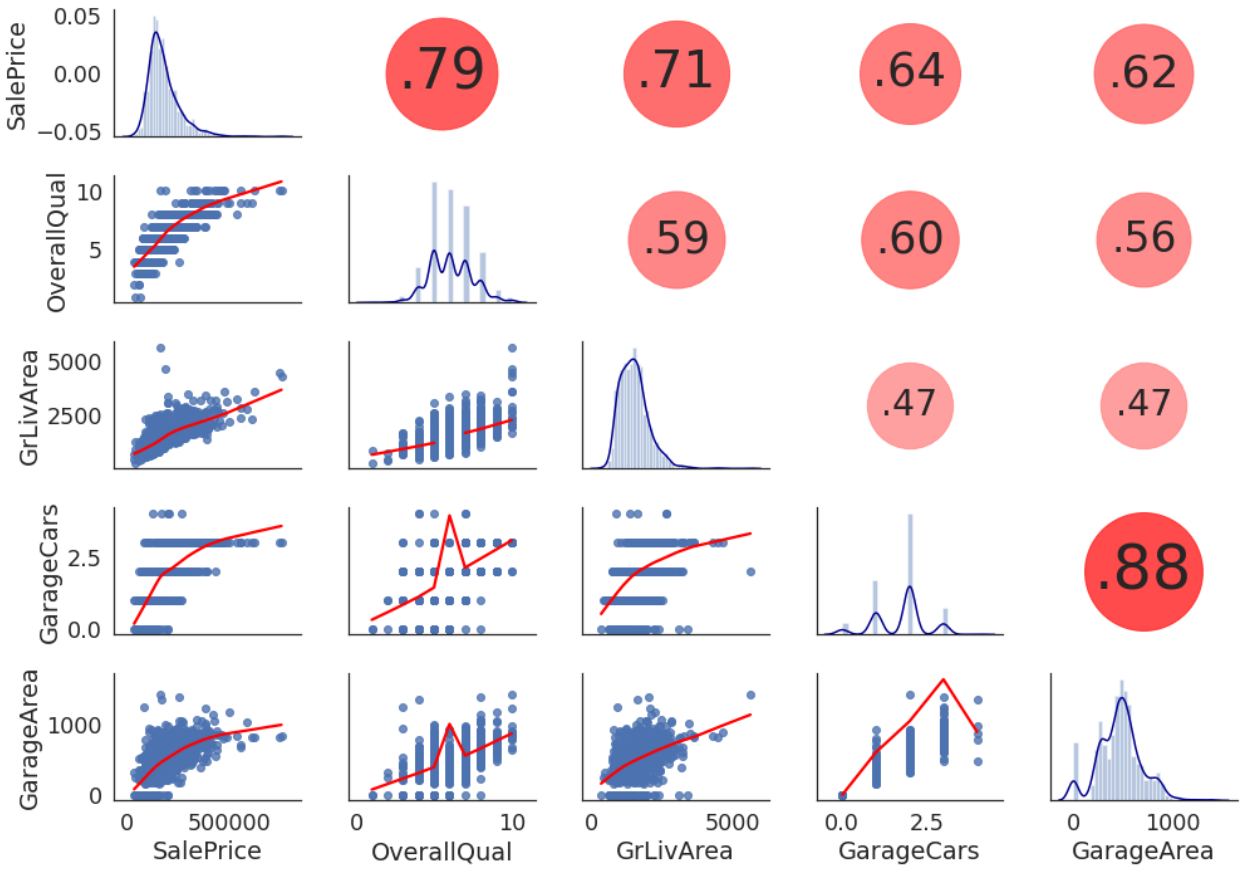
From this histogram (Figure 3), I can see that this data is not normally distributed; the bins are calculated by using the squared method. Since I cannot technically have a negative sales value, I are looking at a graph that is positive skewed. One observation from this distribution plot is that I can confirm that there are no data with weird negative sales values. With a kurtosis of 6.5, this distribution is leptokurtic I large outliers. This would explain the outliers listed around the $800,000 range.

Figure 4 Correlation Heatmap Matrix



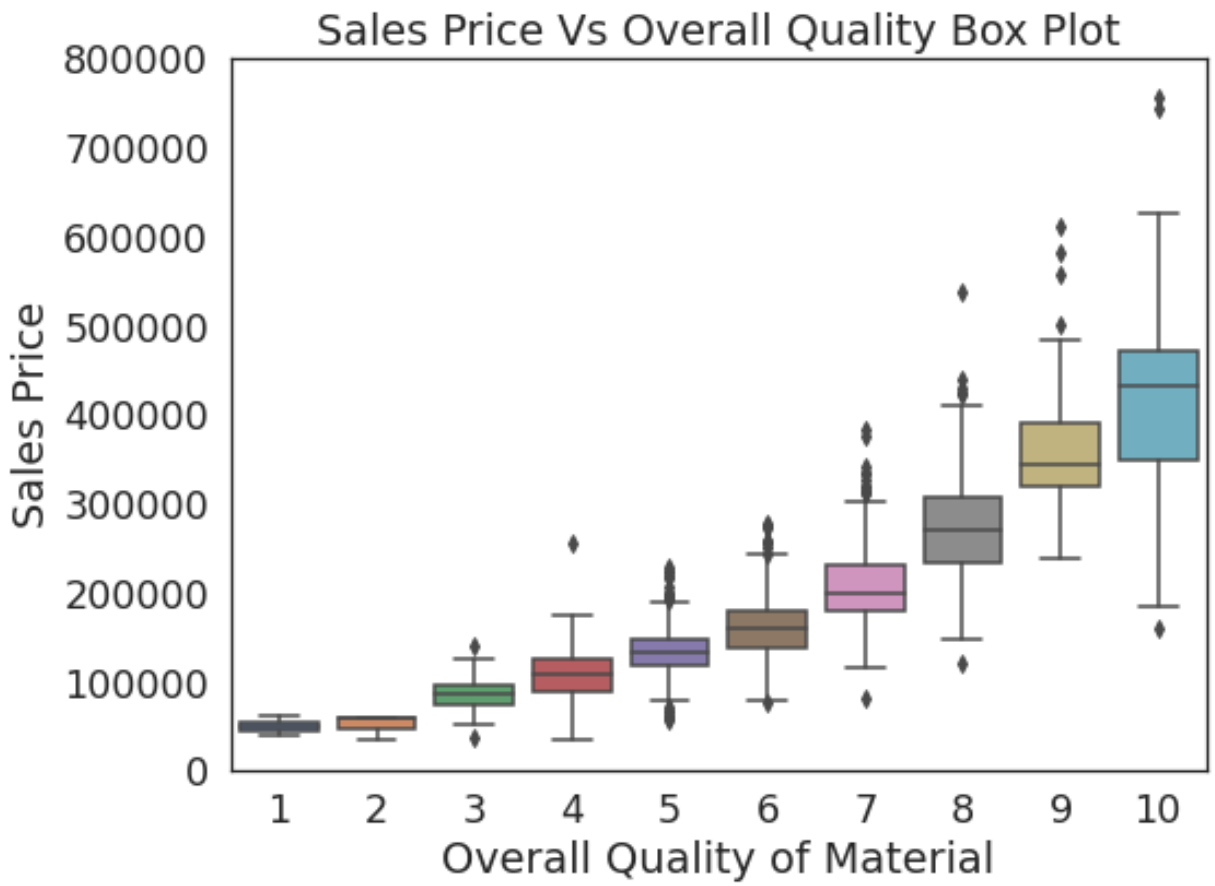
A simple correlation matrix shows the statistical relationship in terms of numbers for the given variables; however, with the heatmap on top of the numeric values for the correlation matrix (Figure 4), one can easily see the positive and negative relationships in a dataset. Since I are observing the sales price, it is important to take note of the sales price within this correlation matrix. Positive relationships of Sales Price include "OverallQual, TotalBsmtSF, 1stFlrSF, GRLivArea, FullBath, GarageCars, and GarageArea."

Figure 5 Zoomed Scatter Correlation Bubble



The zoomed scatter correlation bubble (Figure 5) allows us to see the correlations with their visual regression lines and distributions. By ordering the sales price from most positive correlation of the top five rankings, I are derived with this correlation bubble. From the sales price, I can confirm that there is correlation with the overall quality of the material of the house. Granted with the higher quality, price would range with the material cost. Another observation to note would be the Garage Car (capacity calculated by number of cars) and Garage Area. It makes sense that a garage that can hold more cars would require a larger area. With that in mind, my correlation bubble shows that these variables are also positively correlated with the sales price. It seems that the larger the space and capacity, the more the sales price would increase.

Figure 6 Sales Vs Overall Quality

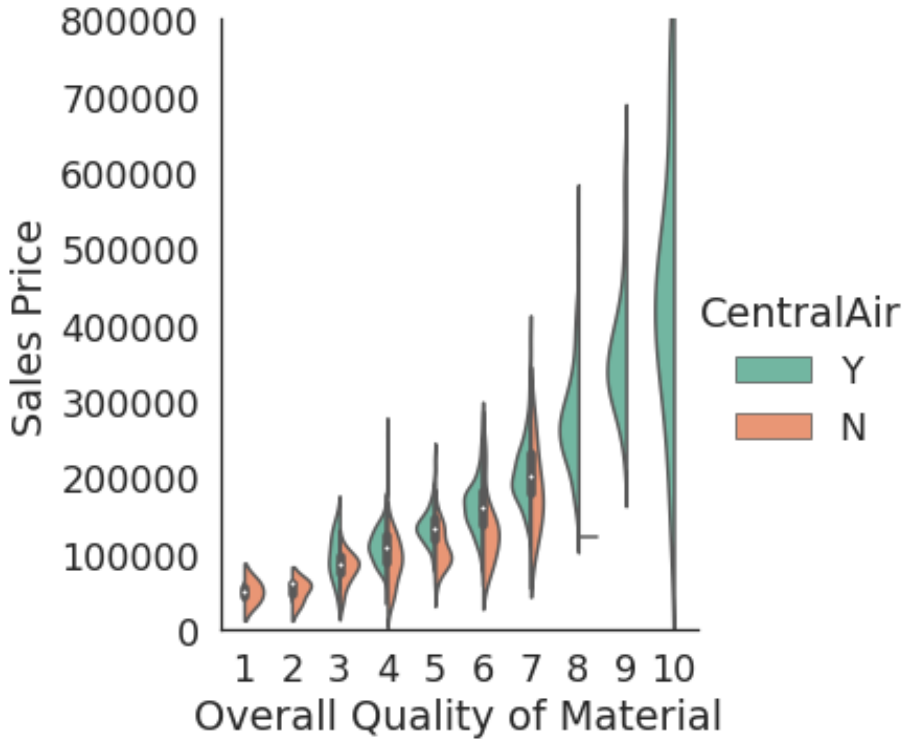


OverallQual: Rates the overall material and finish of the house

10 Very Excellent
9 Excellent
8 Very Good
7 Good
6 Above Average
5 Average
4 Below Average
3 Fair
2 Poor
1 Very Poor

From this boxplot (Figure 6), the mean from each quality shows that there is a clear rising central tendency in sales price. I can confirm that the higher the quality the greater the sales price. I can also make point of the outliers listed in each quality.

Figure 7 Sales Price Vs Overall Quality Multi-Modal Violin Plot



A violin plot helps us understand the distributions of possible separate categories. In my case (Figure 7), I use it to see if central air carries any value to the overall quality of material. I wanted to see the importance of central air with comparison to the overall quality and sales price. It appears that the lower quality homes only have no central air, while the higher quality homes only have central air.

Conclusion

Although sales are impacted by many factors, my analysis shows that sale price of homes in Ames, Iowa are impacted by the overall quality and size of the property. Even though my original hypothesis of the neighborhood location was incorrect, I still managed to conclude with similar results on house size; the overall quality came to a surprise to me. I thought that a house nearby a public or educational facility with great recreational grounds would increase the cost the property due to the living quarters being nearby these varieties of areas. Further investigation for future purposes would include textual analysis of the street addresses and possible listed geographical pinpoints to understand poverty levels of neighborhood locations.

Reflection

Overall this project was very insightful to me since I will be diversifying myself to real estate in the future; thus, selecting the domain of research was not too difficult for me to find. As I have done many projects in regard to scientific research, formulating a simple problem and considering a hypothesis was pretty easy. I believe that if I want to formulate a more complex problem that also includes personally collecting data, then this process would be more difficult. Framing the question to stakeholders and business holders had its own obstacles because one has to understand all the factors that real estate. One challenge during this process was obtaining the data from Kaggle and importing it to Google Collaboration. Considering that I've used python for educational purposes for the past 3 years, I would say that I have an intermediate level of understanding for the code. Implementing Kaggle Dataset into Google Collaboration was more so of myself reading the code and following the instructions. Since I have done several projects for data analysis, I do know what needs to be done to scrub the data and basic EDA. Another challenge would be to be summarizing findings, because knowing how to optimize certain graphs to show depicted results is hard to conceptualize. If my statistics was more refined, I would know exactly what to graph and what to show to my audience, and my findings would be more accurate. In the future I will also better my basic understand of statistics so that I can learn more computer vision techniques for machine learning models.