**Project title:- Effect on Nutrients on Food Ratings**

**Team Name:- Team Pseudo**

|  |  |
| --- | --- |
| Team member Name | SRN |
| Venkata Krishnarjun Vuppala | PES2UG19CS451 |
| SVSC SANTOSH | PES2UG19CS346 |
| KUNTAL GORAI | PES2UG19CS198 |
| S MAHAMMAD AASHEESH | PES2UG19CS342 |

1. **Dataset Name and Description.**

Epicurious - Recipes with Rating and Nutrition.

Over 20k recipes listed by recipe rating, nutritional information and assigned category (sparse).

1. **Problem statement:**
2. To determine how nutritious the food is based on its nutritional value and the important chemicals required by humans.
3. To determine how nutritious the food is based on its nutritional value and the important chemicals required by humans.

To determine how nutritious the food is based on its nutritional value and the important chemicals required by humans.

1. **EDA and Visualization**
   * **How many rows and attributes?**

Number of rows=20.1k

Number of columns=680

* + **How many missing data and outliers?**

**Missing data information:**

|  |  |  |  |
| --- | --- | --- | --- |
| calories | protein | fat | sodium |
| 3749 | 3781 | 3801 | 3750 |

**Outlier information:**

For the categorical variables (ingredients) we had two values 0 and 1 so there were no outliers in them.

For the continuous variables (calories, protein, fat, sodium) we found that there was a total of **2562** outliers.

Outlier finding method used was: **Q1-1.5\*IQR>value or value<=Q3+1.5\*IQR**

* + **Any inconsistent, incomplete, duplicate or incorrect data?**

**Incomplete data:**

We found there are missing data in the continuous variables (calories, protein, fat, sodium). In order to handle those values we replaced it by the mean of that particular column.

**Inconsistent data:**

We found there wasn’t any discrepancies among the column values. Everything was consistent.

**Duplicate data:**

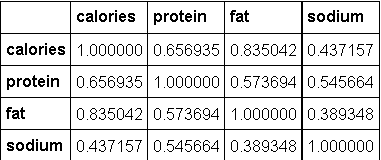
We found 1801 duplicate records in the dataset. In order to handle the duplicity we just dropped the duplicated records.

**Incorrect data:**

We found there were 2562 outliers in the dataset. In order to handle it we dropped those records.

* + **Are the variables correlated to each other?**

To find correlation among columns we created a correlation matrix among the continuous columns (using the corr() method).



From the above matrix we found that there was a high correlation between calories and fat (0.835) and protein and calories (0.656).This might cause a problem of multicollinearity.

* + **Are any of the pre-processing techniques needed: dimensionality reduction, range transformation, standardization, etc.?**

**Standardization:**

As the categorical variables were dichotomous i.e. values were either 0 or 1. There wasn’t any requirement of standardization.

But for the continuous variable (sodium, fat, protein, calories) the range of variables had a lot of variation so we performed a standardization of these variables. The method used for standardization was StandardScaler(). This will help for our further model predictions.

**Range Transformation:**

As our predictor rating was a continuous variable we introduced a new column called rating label which performed a range of values to a given label transformation (concept hierarchy).

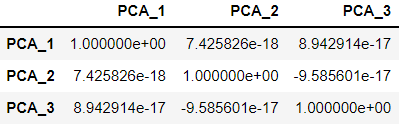
So our rating labels was done as follows:

* 1. ->Very Bad 1-2->Bad 2-3->Average 3-4->Good 4-5->Excellent

This new column helped us in visualization of our data.

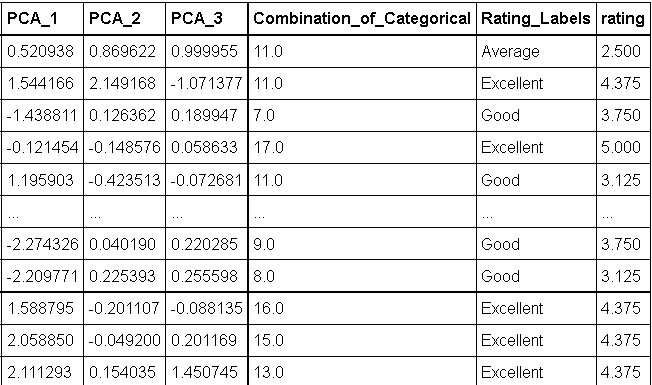
**Dimensionality reduction:**

As from the previous question we found there was a high correlation in some pair of variables. So we performed a PCA transformation to 3 components.



From the above matrix we see that there is very low correlation among them.

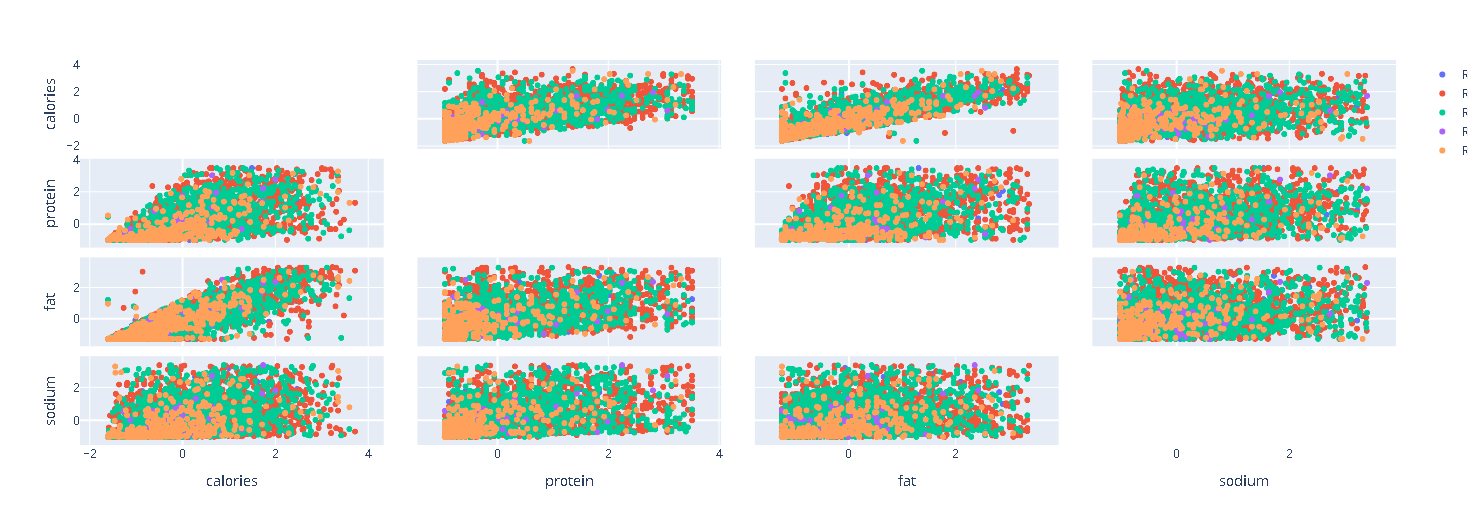
In the dataset we noticed there were around 674 categorical variables. As handling with these columns is a tedious task. So we performed a one hot encoding technique and reduced 674 column to a single column which is sum of all categorical variable.



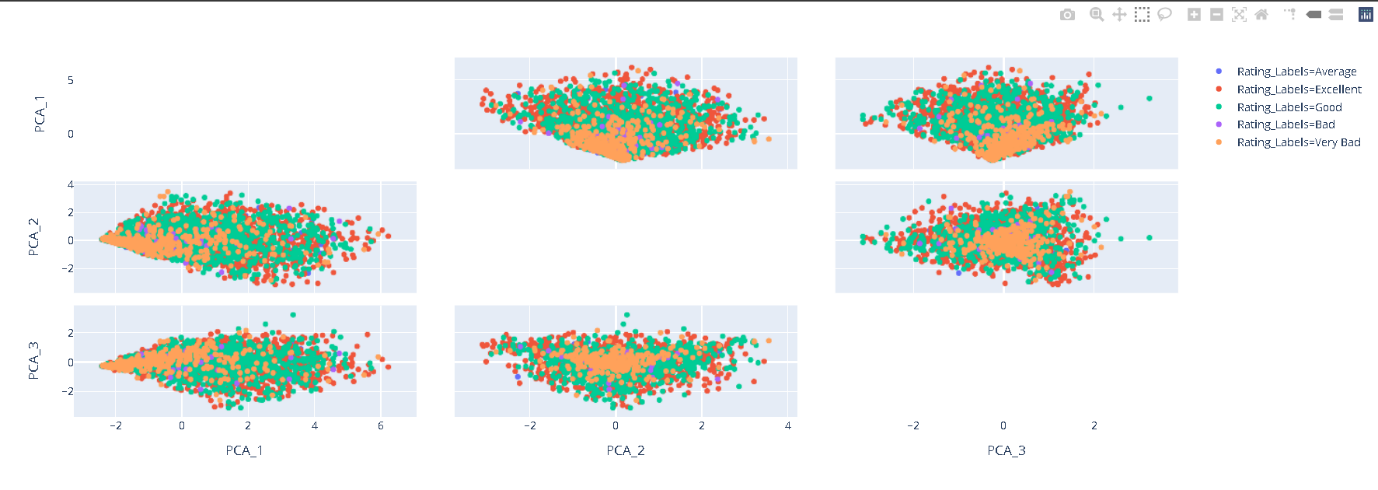
So by using various dimensionality reduction strategies, range transformation we reduced 680 columns to just 6 valuable columns for our further data analysis.

* + **Does PCA help visualize the data? Do we get any insights from histograms/ bar charts/ line plots, etc.?**

We performed a scatter matrix plot of the continuous variable before and after applying pca.



Before applying PCA

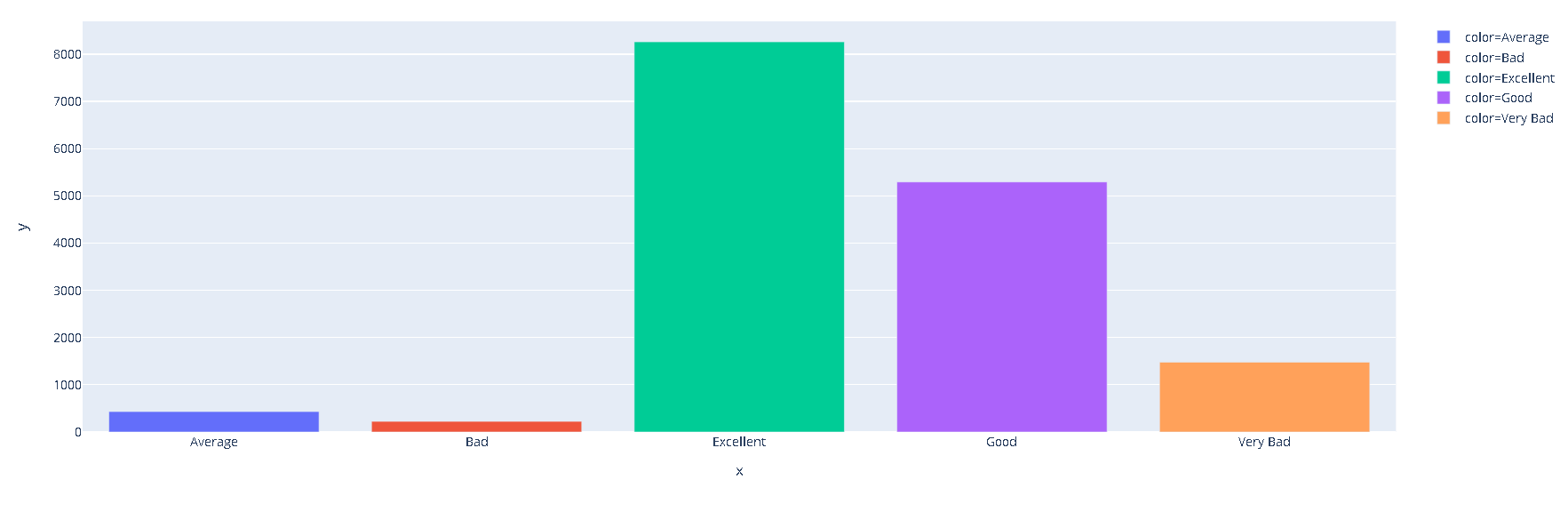


After applying PCA

We see that the plot before pca was randomly distributed making it tough to predict the outcomes. But when we apply pca on these variables we see there is a pattern that emerges which helps us in distinguishing the different rating labels. PCA does help us to visualize the data.

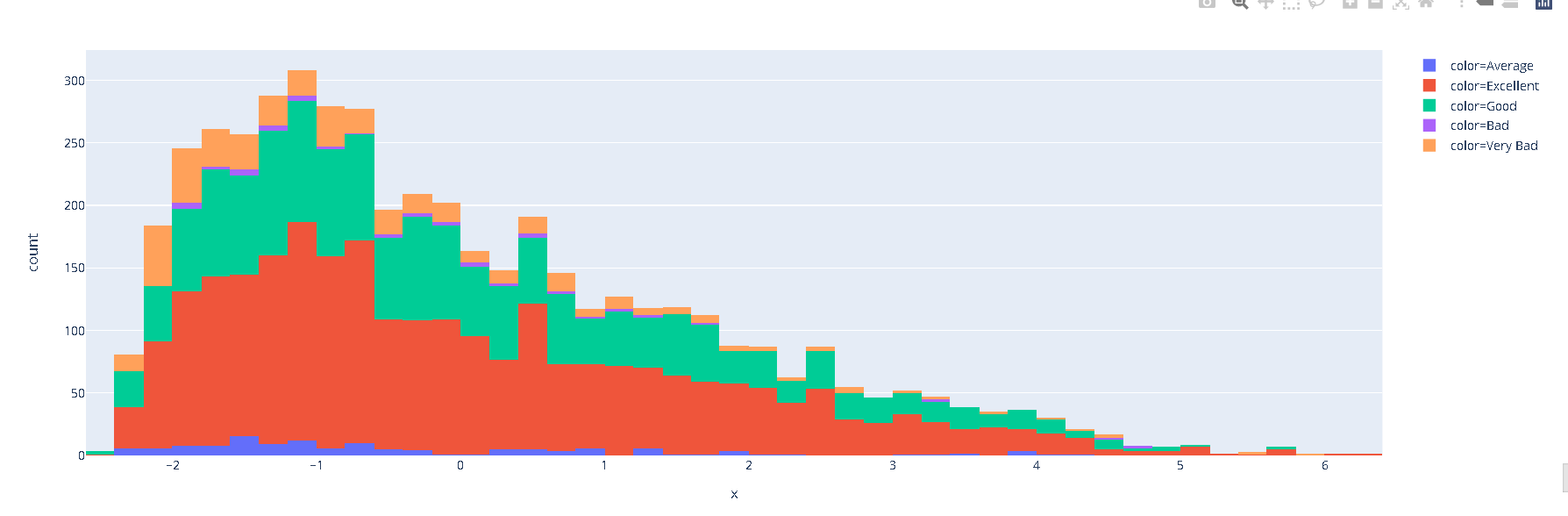
**Insights based on various visualization techniques**:

**Histogram on rating labels:**



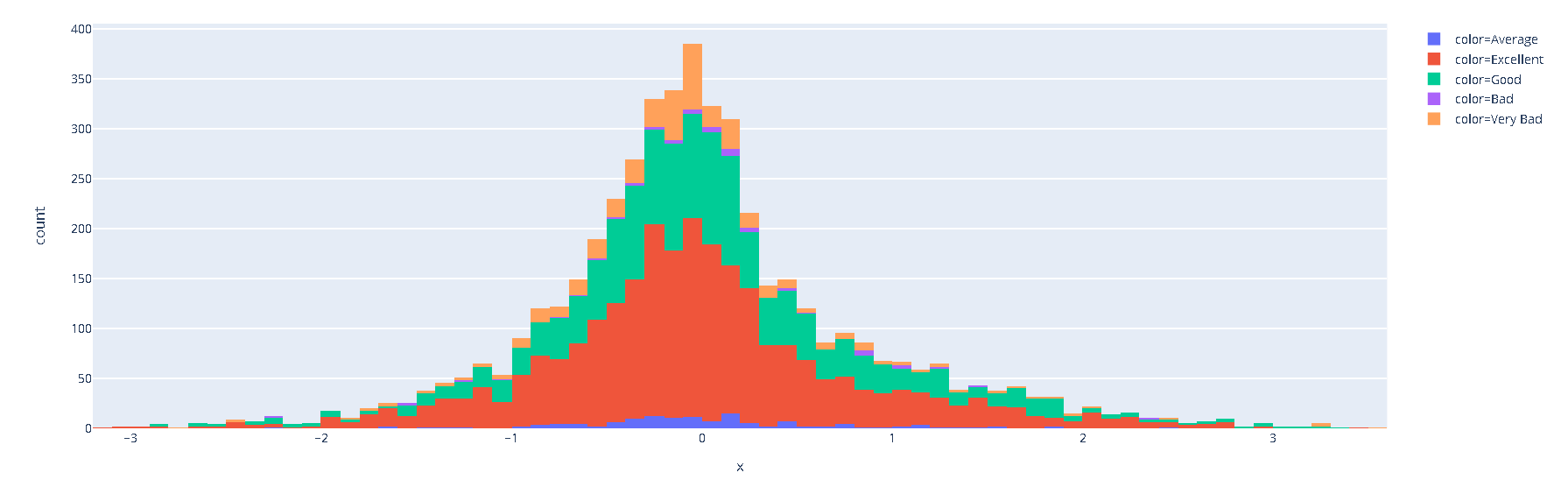
From the histogram we see there is a highest proportion of excellent rating and then good rating. It tells us that there is significant amount 4-5 ratings and 3-4 ratings in our dataset.

We have to make sure that our data can distinguish these two labels more accurately than the others.

**Histogram on the PCA components**

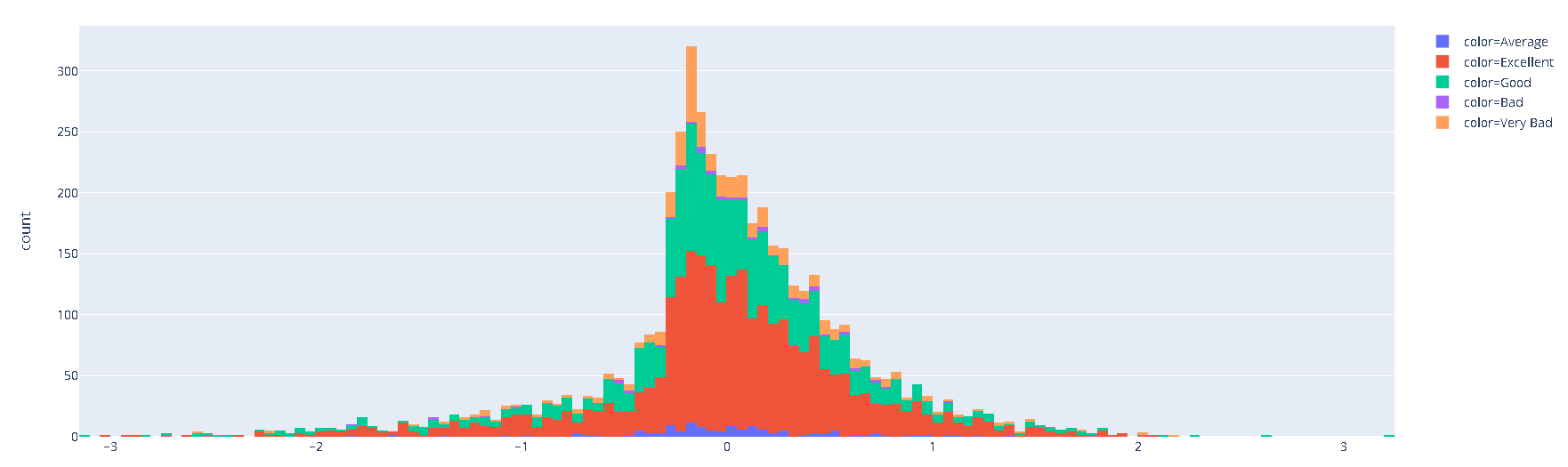
**PCA\_1**

The above histogram has more points towards the left compared to right showing a partial left skewed distribution. This give us an insight that our PCA1 near to range (-2, 1) have more excellent rating’s.



**PCA\_2**

The above is not a skewed distribution. This distribution is perfectly normalized This give us an insight that our PCA2 near to range (-1, 1) have more excellent rating’s.



**PCA\_3**

The above is not perfectly normalized but there is high skewness in the left side of data. This give us an insight that our PCA3 near to range (0, 1) have an excellent rating’s.

1. **Link for google sheet:**

<https://docs.google.com/spreadsheets/d/1njYwrPHOdCm2SaPskVaw7vcp7c6SRrzohvNdK6nd0zc/edit?usp=sharing>

1. **Literature Survey (Summarize):**

**The primary objective of the model is to determine a rating, or score based on the given independent variables or attribute values. The other questions that problem will either solve or give more insight are:**

* **How much nutrition is required for a particular food on an average for having a well-balanced nutritional and healthy value?**
* **What proportion of nutrients is unhealthy i.e., what is the conjunction of attribute values that determines the score of the food which in-turn determines whether the food is healthy or not?**
* **Should there be a balanced amount of biochemicals, substances that the human body must intake for a balanced diet?**

**The dataset obtained has 20000+ records with a good number of attributes and is well spread with respect to nutrients values. However, the problems faced with the dataset are:**

**There are several outliers for some attributes which may seem irrelevant but may have some value with respect to the model. This serves as a problem since outliers may be misleading.**

**There are a lot of unimportant attributes present in the dataset which provide little to no value with respect to the ratings or score given to the model.**

**The determination of which attributes are important is difficult due to the number of attributes present.**

**Despite all this, these problems can be solved with the help of the following methods:**

* **PCA (Principal Component Analysis).**
* **Normalization.**
* **Removing outliers which provide little to no value to the model.**

1. **Your Plan:**

**The plan is to use Multiple Linear Regression where it takes in a collection of attributes and computes a value(rating). The dataset is being divided into training, validation and test parts which helps in determining the accuracy, precision, confusion matrix, etc.**

1. **References**

* <https://www.datacamp.com/>
* <https://machinelearningmastery.com/>
* <https://globaljournals.org/>
* <https://www.cse.iitb.ac.in/~uday/soft-copies/writing-a-good-paper.pdf>

1. **Link to EDA Analysis:**

https://github.com/Recipe-reccomendation-System/EDA-Analysis