vmw VMware codehouse India 2021

Web Intelligence

TEAM -7

BOOLEAN AUTOCRATS

TEAM MEMBERS

MANGIPUDI SRUTHI

PALAK KOTWANI

KEERTHANA S

MAHATHI MUDDHEY

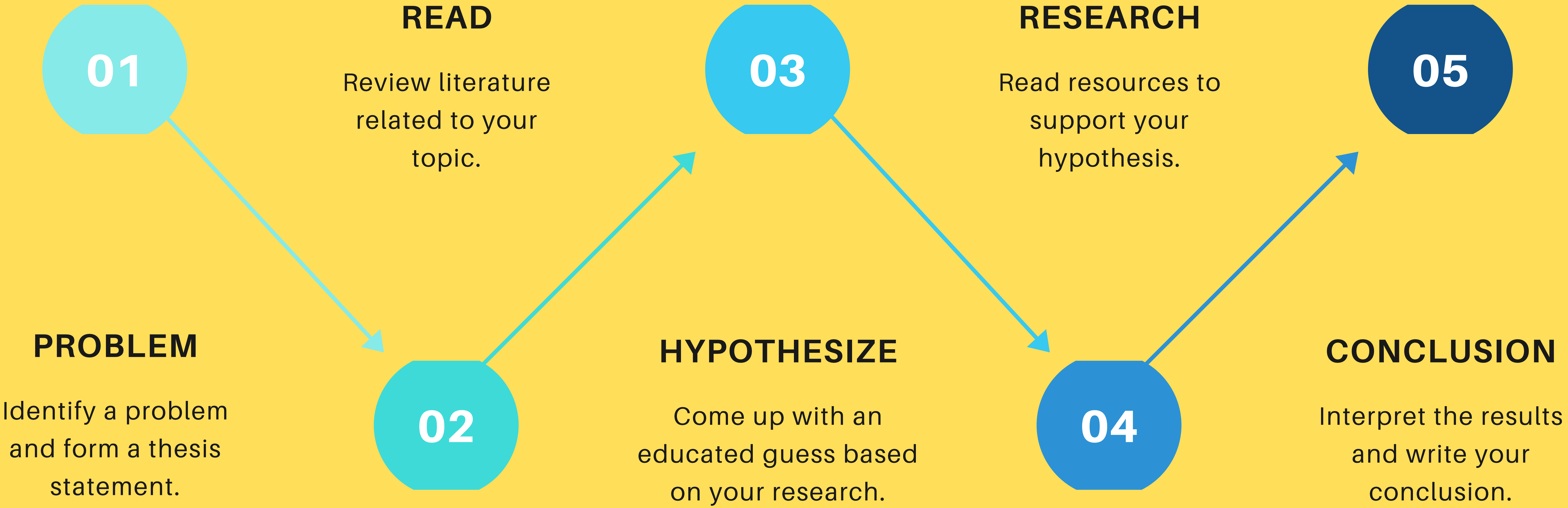
Key points

- Introduction
- Problem Statement
- Web Scraping
- Data Cleaning
- Data Analysis
- Data Visualization
- Conclusions
- Further Improvements

Introduction

- Web Intelligence helps you enable all business users to understand trends and root causes with easy-to-use tools for ad hoc queries, reporting, and analysis in a Web environment.

Typical Project Analysis Approach Followed



01

PROBLEM

Identify a problem and form a thesis statement.

02

READ

Review literature related to your topic.

03

HYPOTHESIZE

Come up with an educated guess based on your research.

04

RESEARCH

Read resources to support your hypothesis.

05

CONCLUSION

Interpret the results and write your conclusion.

Problem Statement

- Primary Analysis of VMware SD-WAN with Web Scraping NLP (Natural Language Processing) and Sentiment Analysis.
- Apply the same analysis methods to Cisco SD-WAN and compare the results by doing a cumulative study of the reviews to concur the product success rate.

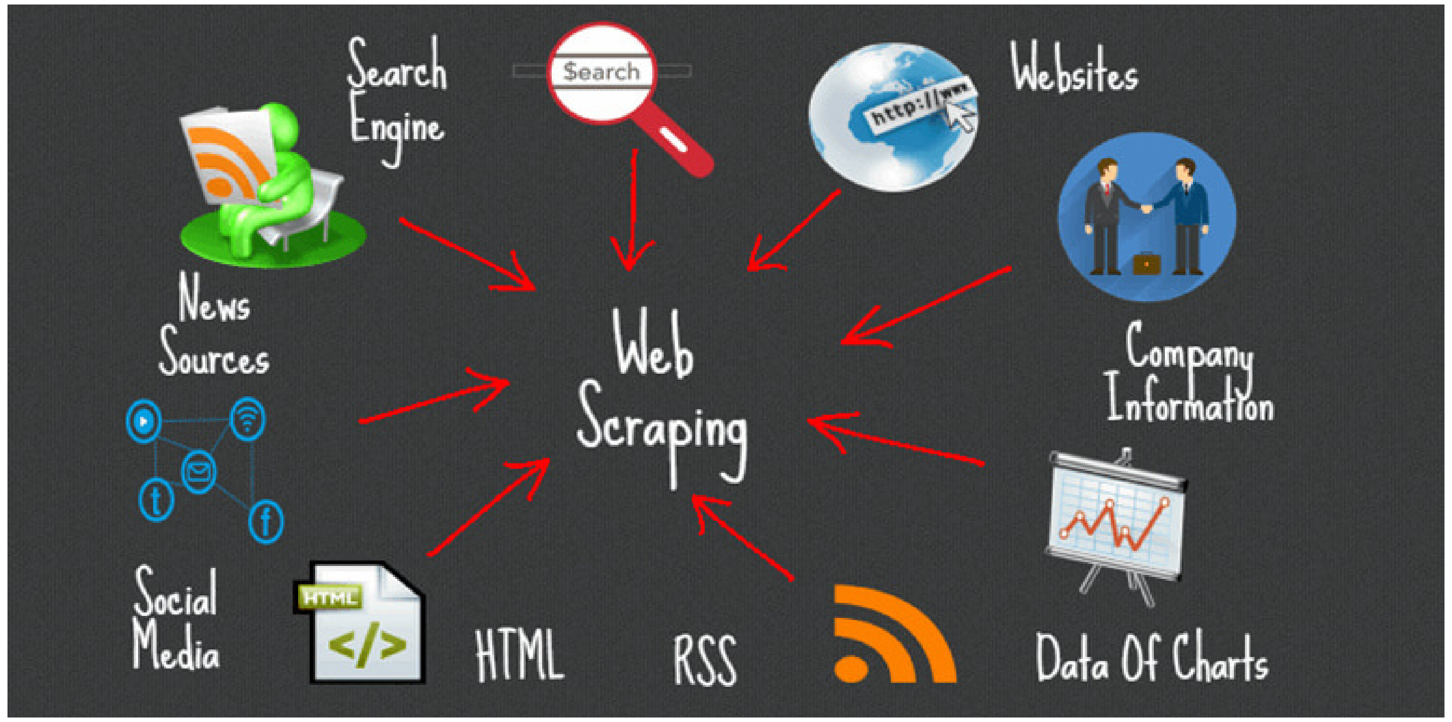
Web Scraping

- Web scraping is the process of using bots to extract content and data from a website
- Instant Data Scraper is the Tool used for Data Scraping Part. It provides quick and accurate data which could be easily converted to a csv file .
- For Web scrapping we considered the following two websites to compare and collecting the reviews of the users for the Cisco SD -WAN and VMware SD-WAN

- https://www.gartner.com/reviews/market/wan-edge-infrastructure/vendor/cisco/product/cisco-sd-wan
- https://www.gartner.com/reviews/market/wan-edge-infrastructure/vendor/vmware/product/vmware-sd-wan

Web Scraping

COMPARED PRODUCT DETAILS



VMware SD-WAN

Data set used : CSV File (VMware) : https://drive.google.com/file/d/1sTt6et4z7ELyB84ZzygNxzxLNXQ02cLi/view

Cisco SD-WAN

Data set used : CSV FIle (CISCO) : https://drive.google.com/file/d/1PKiYb070BYjW73oKpEj5MYXVlls-M9yJ/view?usp=sharing

- The Data is an Integral part of analysis and it needs to be accurate and Free from the Nil values so that it doesn’t affects the Results.So here is the next step tp deal with the cleaning /preprocessing part of the data

Data Cleaning

- Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.
- The data which we got had inessential columns and corresponding data so during this step we dropped them from the original dataframes
- Also as we had no mathematical values in our dataset so there wasn’t any need to handle nil values or do any categorical analysis.

Data Cleaning

- The complete review is appended to the review headline.
- The punctuations and stop words are removed.
- The reviews are tokenized.
- We create Parts of Speech Tags for each token.
- Sentiment Analysis is carried out on the tokens to obtain mathematical results that are better fit for visualization.

Data Analysis

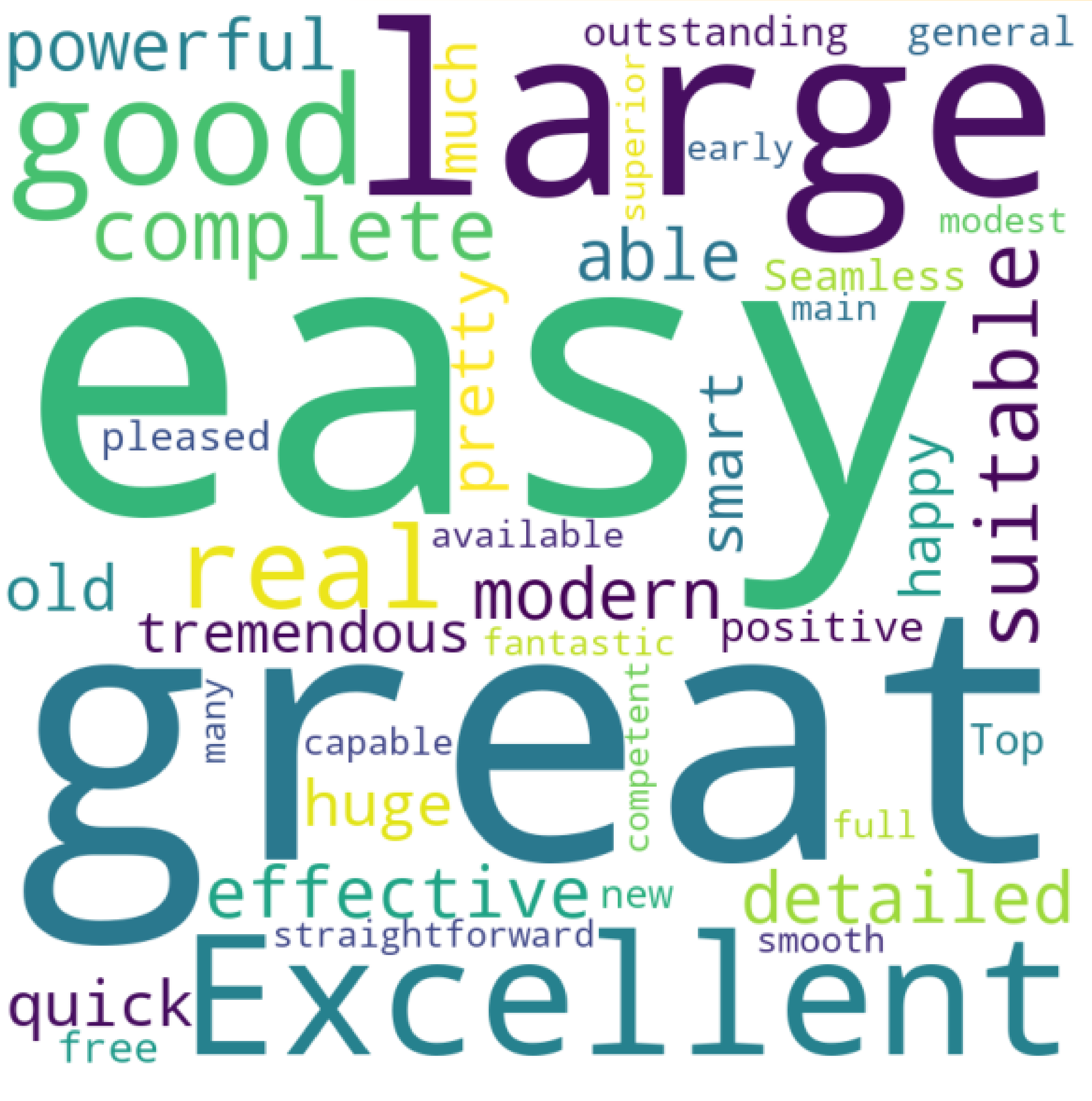
- Once Raw data is cleaned - After removing punctuations, stopwords, We are left with only required information which is so much easier to do analysis on.
- We use Dictionaries in python to accumulate similar data together and plot them using matplotlib to visualize the same.
- We also device a simple mathematical model to find the success rate of a products and compare those two. Although the model is very subjective, It helps us grasp the results better.

Sentiment Analysis

- We use TextBlob package to carry out the sentiment analysis of the reviews. As TextBlob itself has only 60% of working efficiency, We clean the data and use only the relevant tokens with appropriate tags to carry out the sentiment analysis which gives us better results.
- We obtain the polarity and subjectivity of the reviews where polarity refers to the tone of the review- positive being good and negative being bad. The subjectivity refers to the degree of personal opinion.
- Besides plotting We use these values later again in our mathematical model to obtain the success rate of the product.

How users see VMware SD-WAN

With the cleaned tokenized data, we pick out the adjectives using POS tags to see what users think about VMware WAN in a word cloud

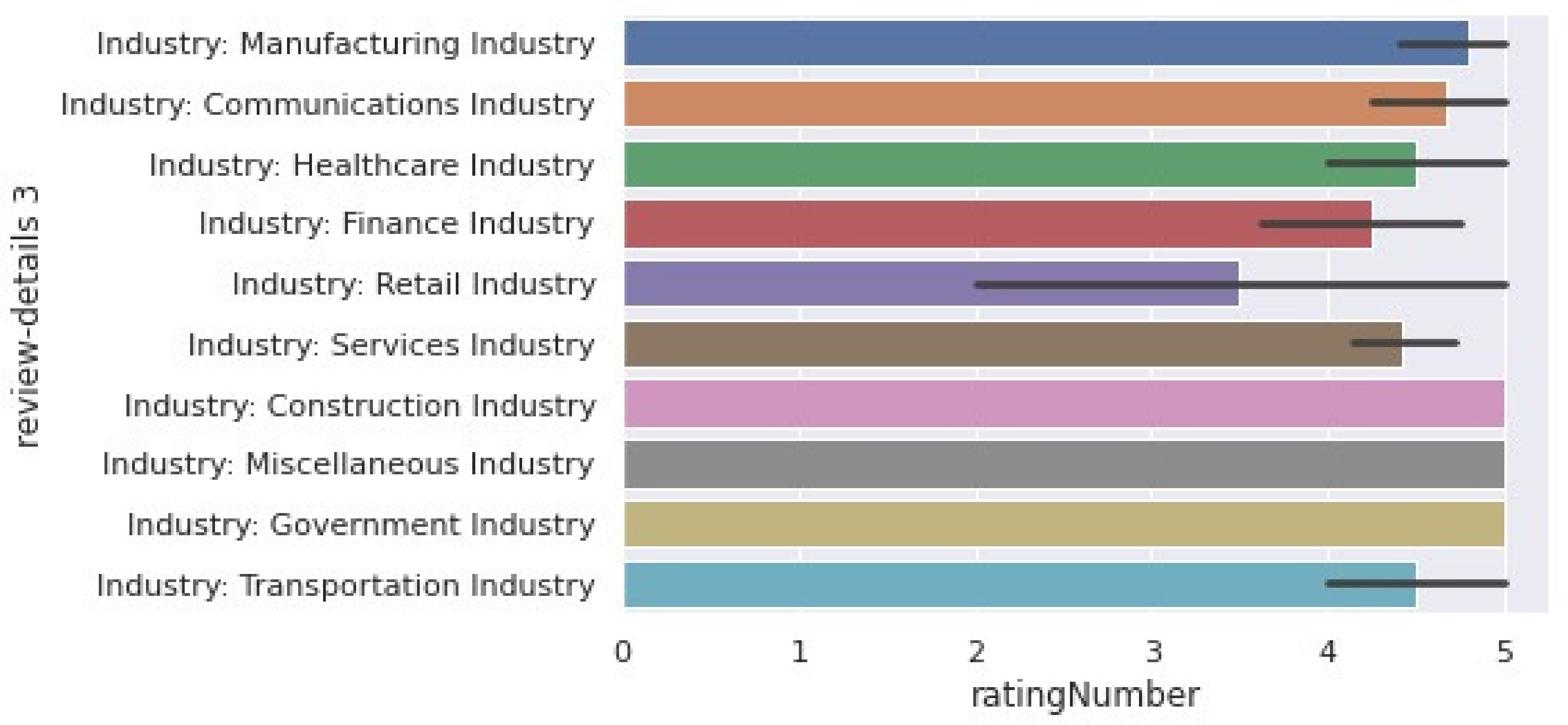


Data Visualization

- Post Web Scraping, data cleaning and data analysis, the process involves data visualisation which gives us a clear idea of what the information means by giving it visual context through maps or graphs.
- In our project we have used the seaborn and the matplotlib libraries in order to carry out the data visualisation
- The following data has been visualised with the help of these libraries

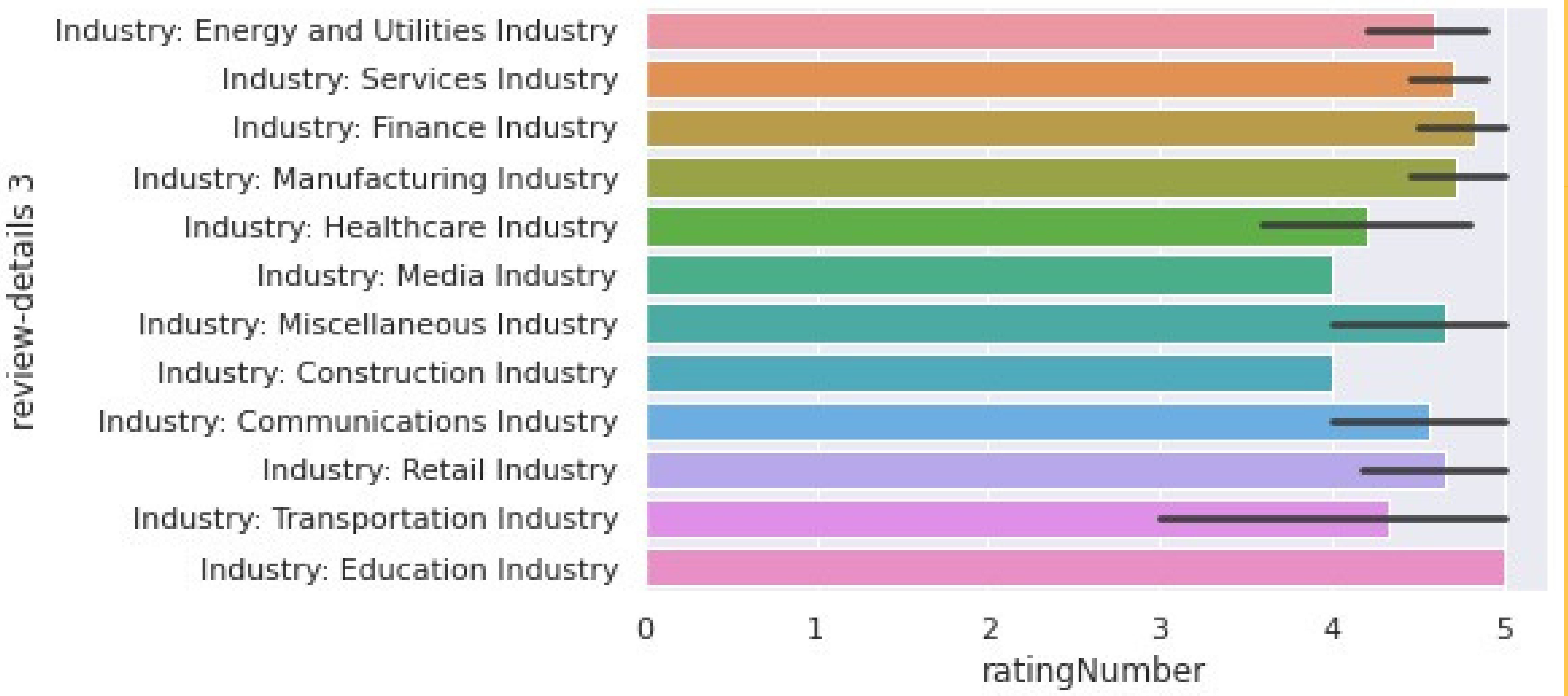
Data Visualization

- Review ratings of VMware products given by various industries



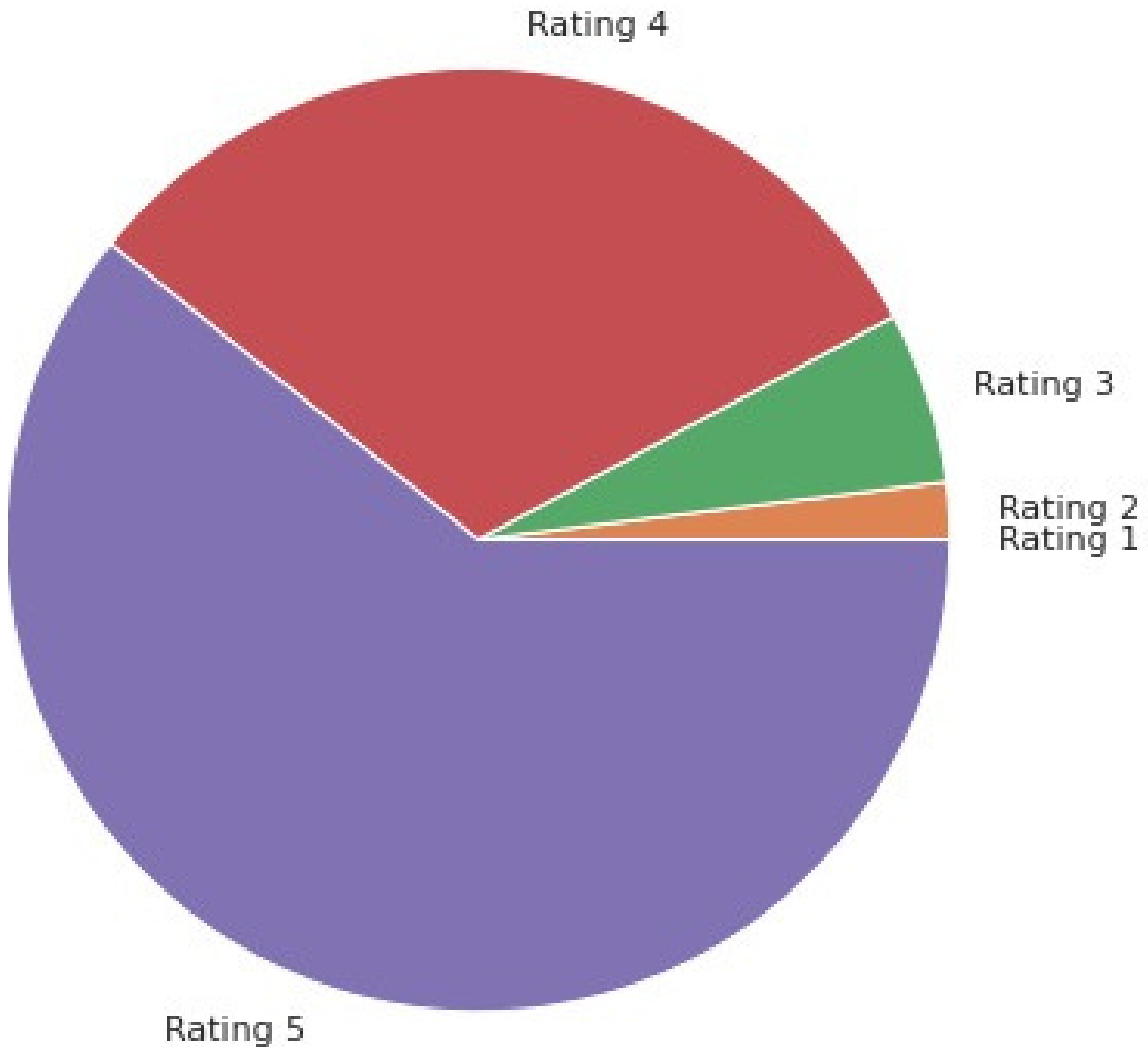
- The ratings here indicate that the government and the construction industries are the most satisfied consumers of the product

- Visualisation of the review ratings of other competing companies also helps us have better insights into understanding where the company stands in terms of customer satisfaction. Therefore in our project , the ratings and the reviews of the Cisco company products have been used for comparison.
- The following visualizes the review rating of the Cisco product given by the users from various industries

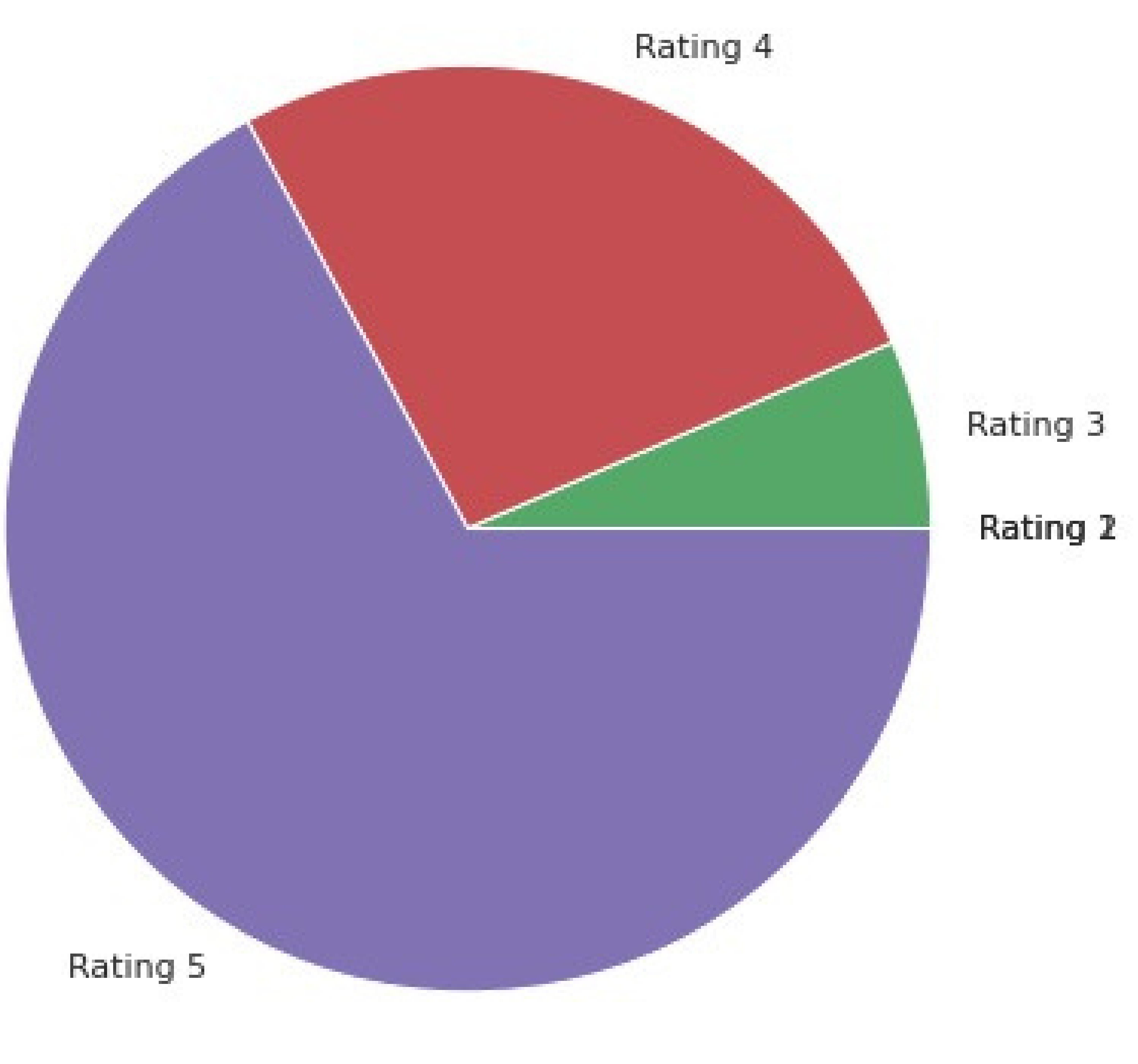


- Visualizing the ratings of the VMware products vs Cisco products

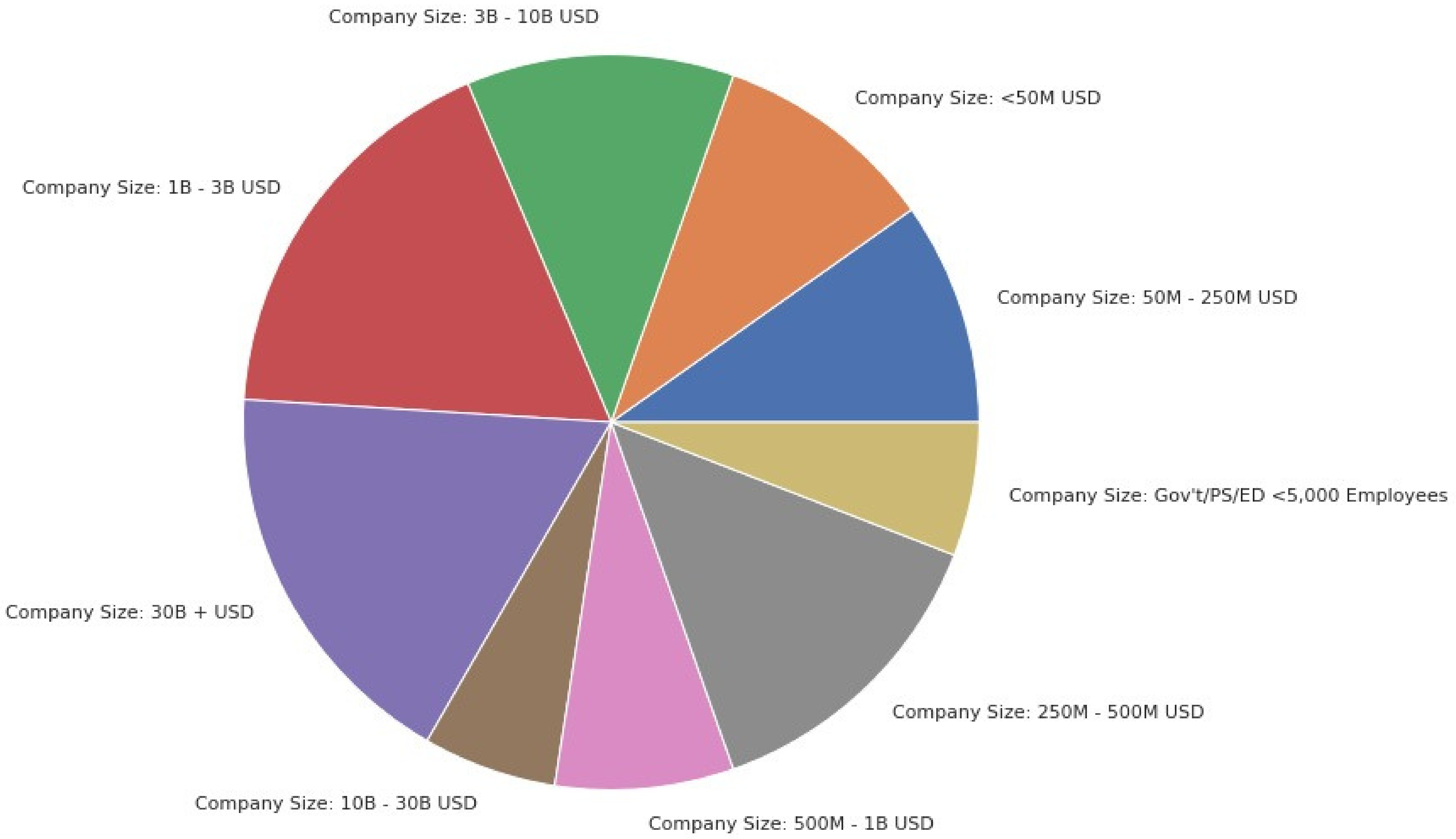
VMware



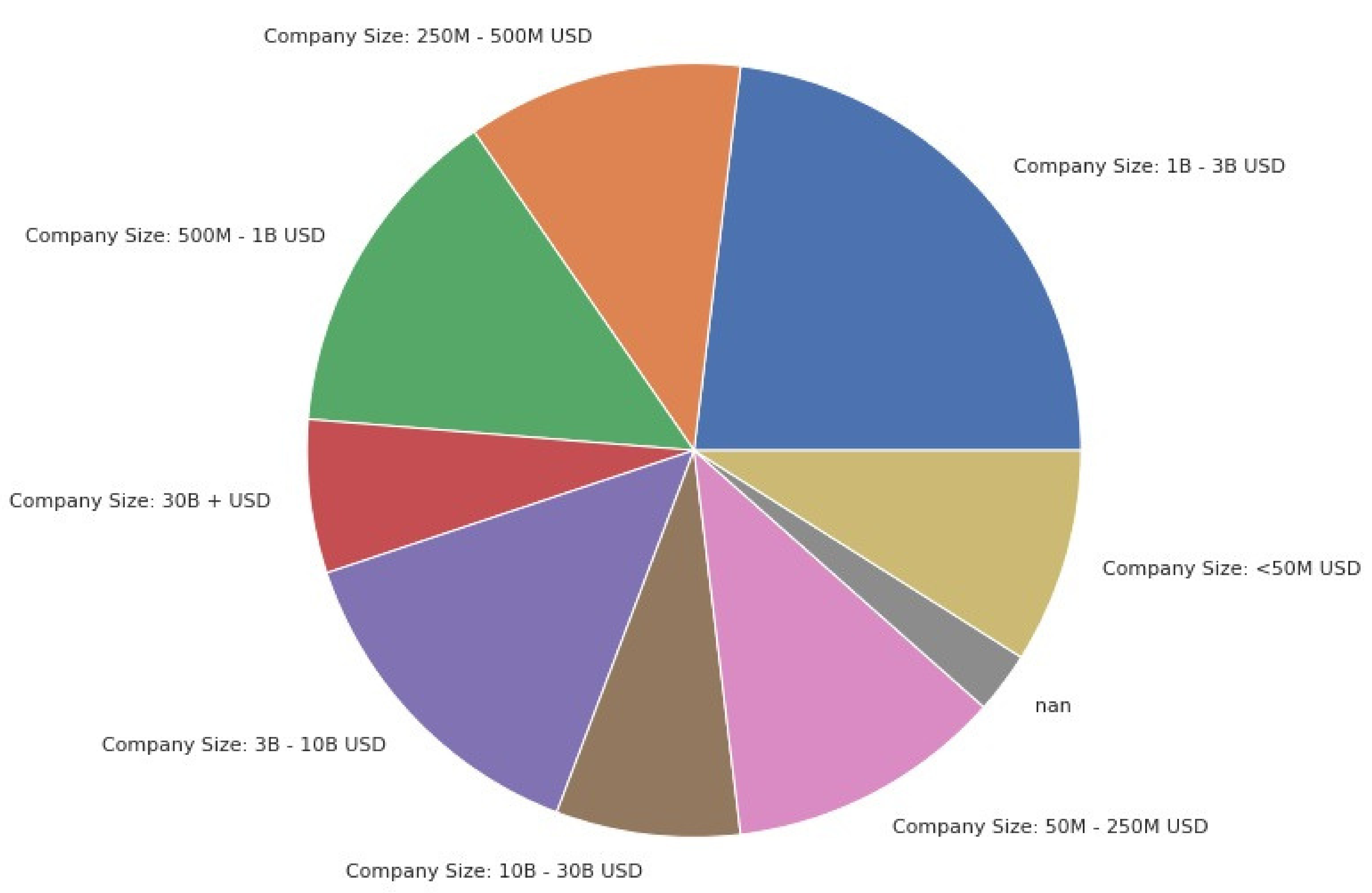
Cisco



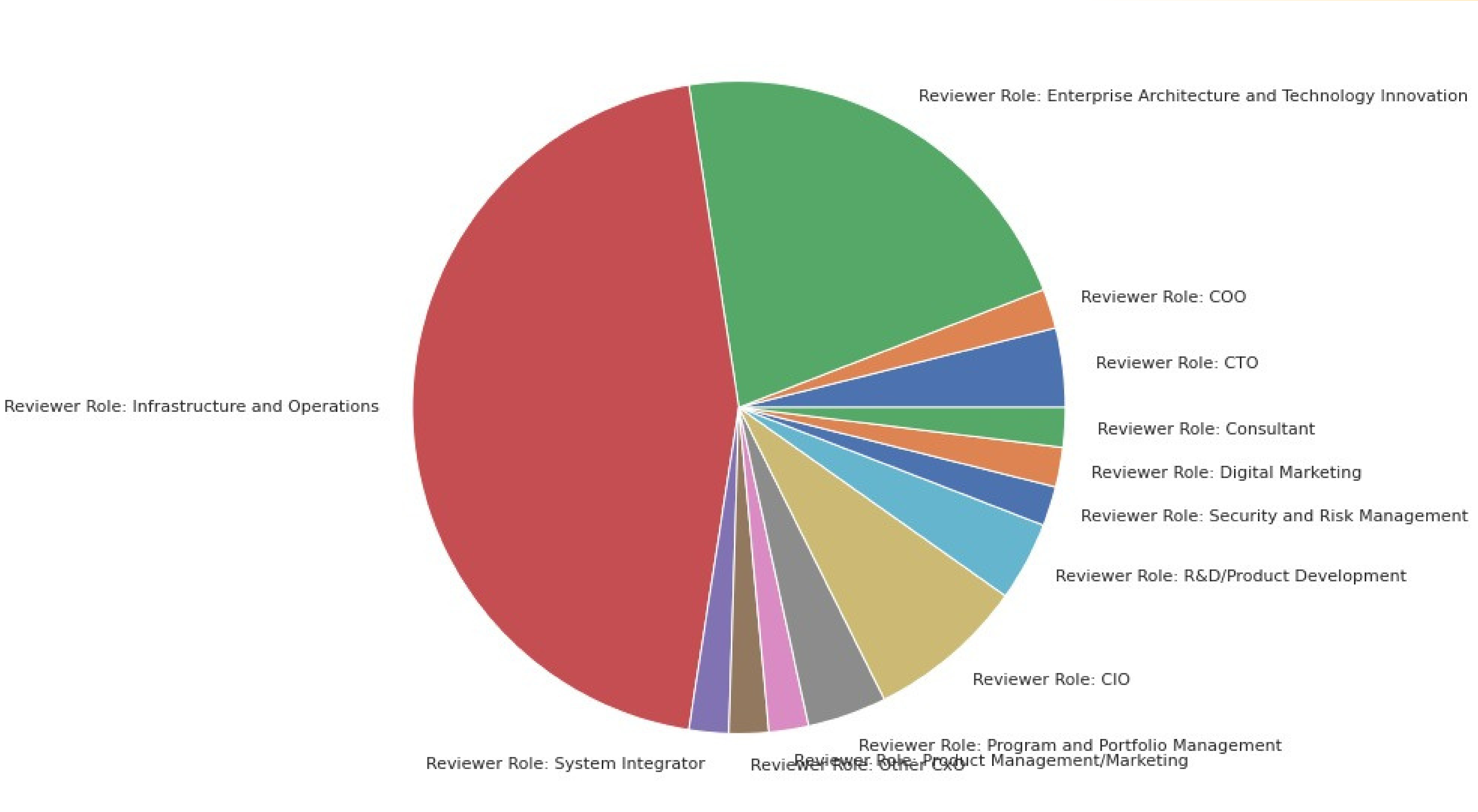
- The following is the visualisation of the net worth of the various companies using the VMware product



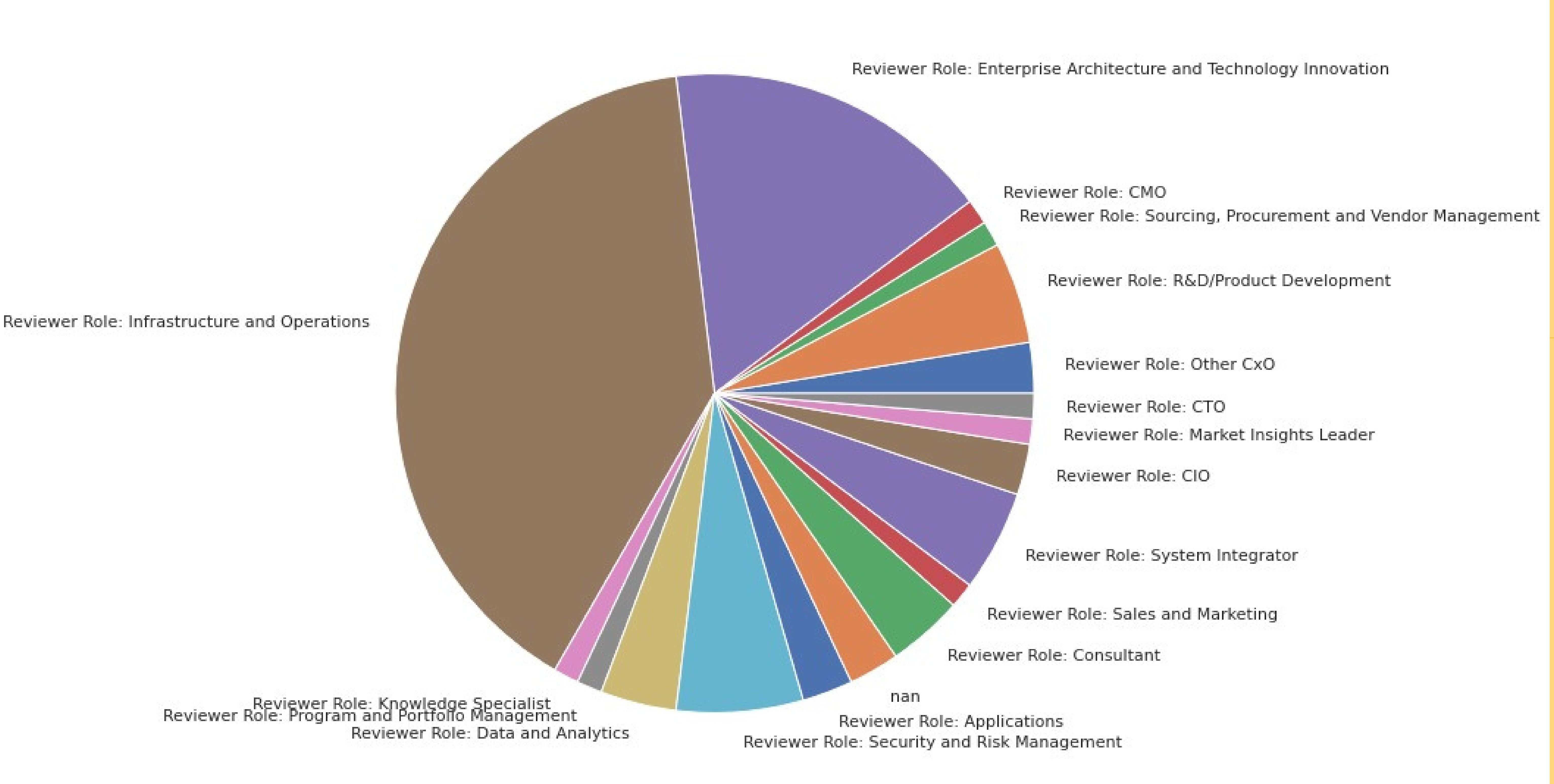
- The following is the visualisation of the net worth of the various companies using the Cisco product



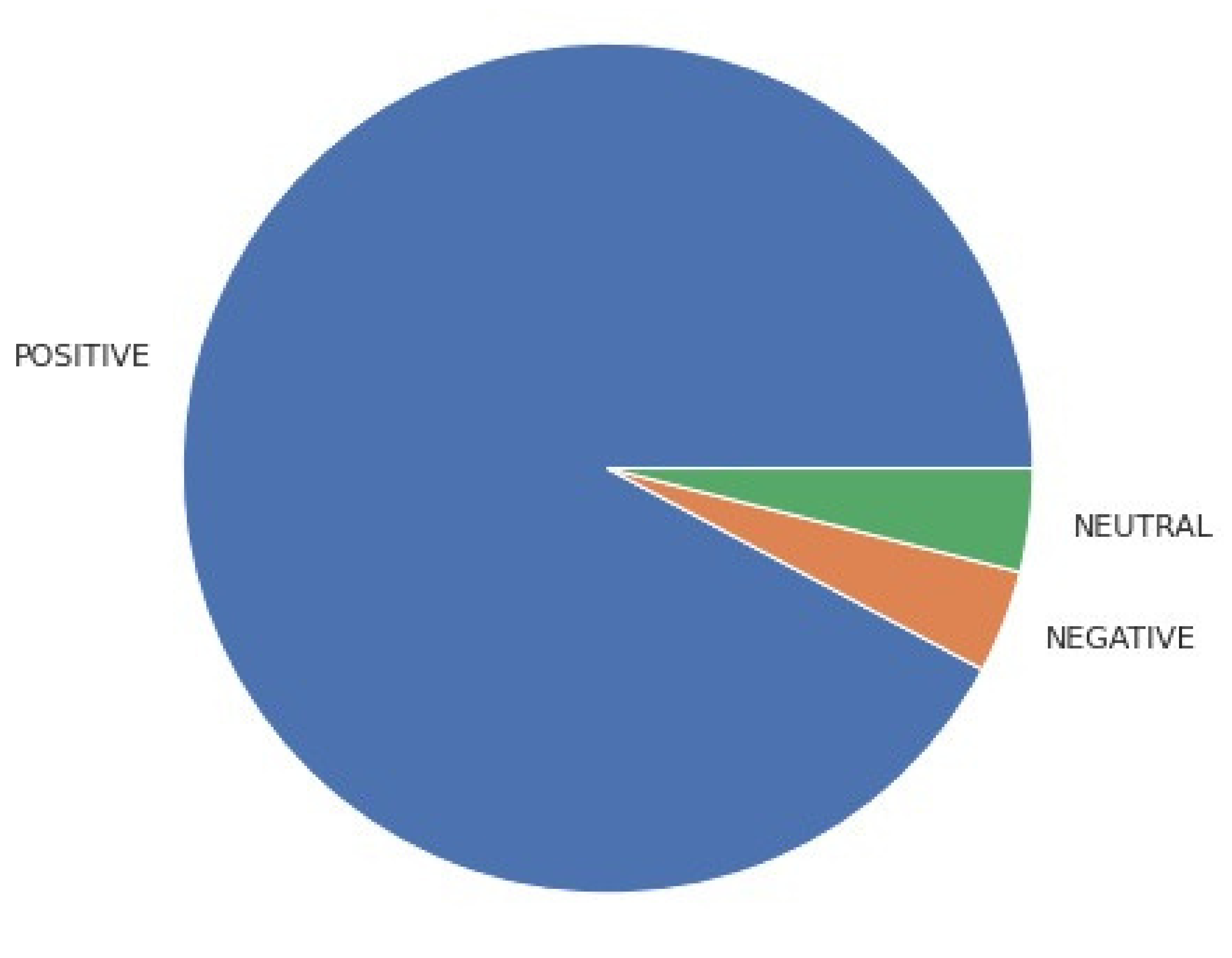
- The participation of the users of various designations in the reviewing of the VMware product is visualized in the form of the pie chart below

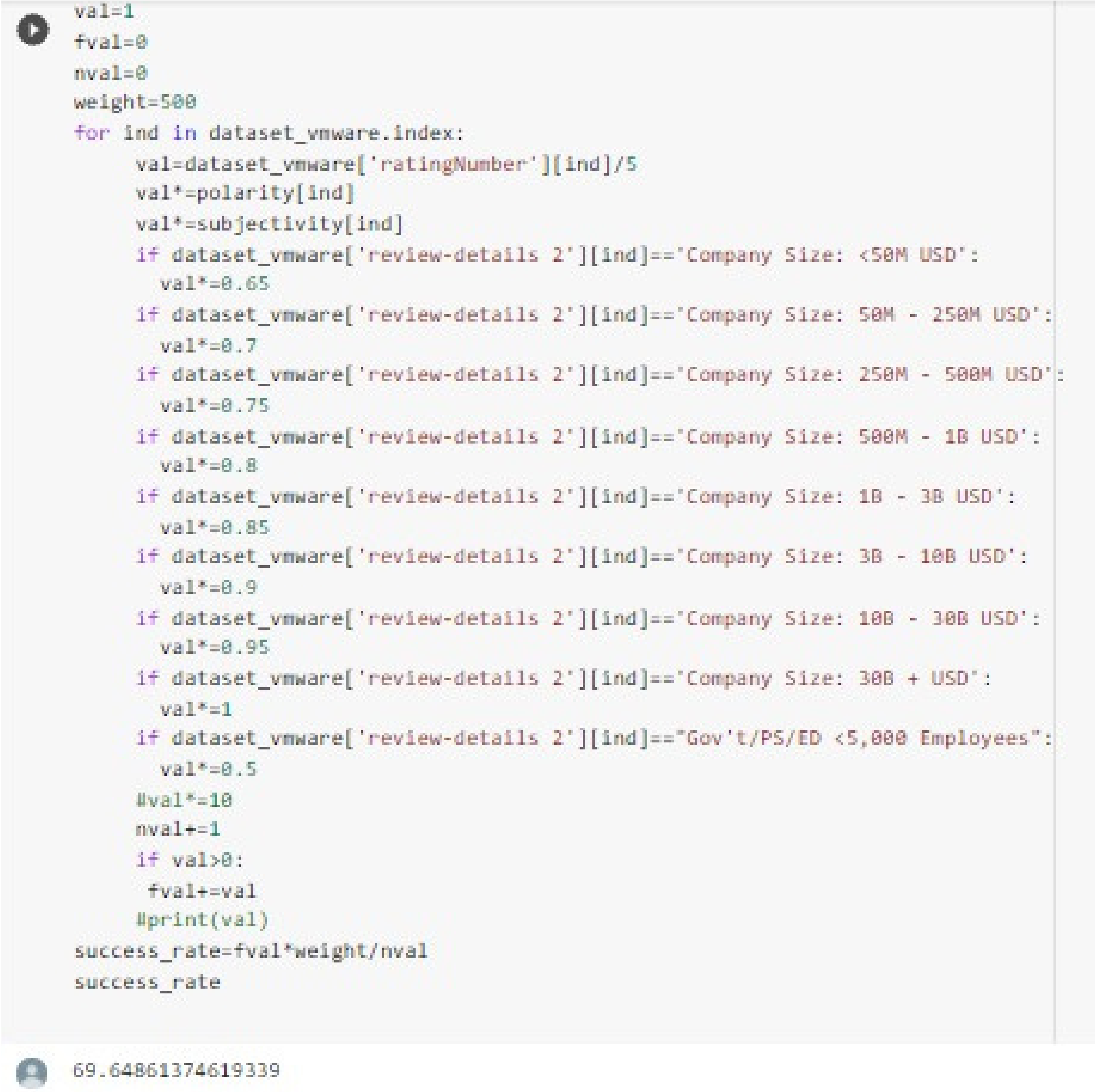


- The participation of the users of various designations in the reviewing of the Cisco product is visualized in the form of the pie chart below



- Visualization of the sentiment of the reviews





```
val=1
fval=0
nval=0
weight=500
for ind in dataset_vmware.index:
    val=dataset_vmware['ratingNumber'][ind]/5
    val*=polarity[ind]
    val*=subjectivity[ind]
    if dataset_vmware['review-details 2'][ind]=='Company Size: <50M USD':
      val*=0.65
    if dataset_vmware['review-details 2'][ind]=='Company Size: 50M - 250M USD':
      val*=0.7
    if dataset_vmware['review-details 2'][ind]=='Company Size: 250M - 500M USD':
      val*=0.75
    if dataset_vmware['review-details 2'][ind]=='Company Size: 500M - 1B USD':
      val*=0.8
    if dataset_vmware['review-details 2'][ind]=='Company Size: 1B - 3B USD':
      val*=0.85
    if dataset_vmware['review-details 2'][ind]=='Company Size: 3B - 10B USD':
      val*=0.9
    if dataset_vmware['review-details 2'][ind]=='Company Size: 10B - 30B USD':
      val*=0.95
    if dataset_vmware['review-details 2'][ind]=='Company Size: 30B + USD':
      val*=1
    if dataset_vmware['review-details 2'][ind]=="Gov't/PS/ED <5,000 Employees":
      val*=0.5
    #val*=10
    nval+=1
    if val>0:
     fval+=val
    #print(val)
success_rate=fval*weight/nval
success_rate
```

```
69.64861374619339
```

- On carrying out the success rate analysis on each of the company datasets, the following success rates have been established
- The VMware dataset indicates an overall success rate of 69 .64 percentage while that of Cisco is 52.87 percent.

Conclusions

- We can know which industry consumers contributes more/less for a product, which industrial area necessaries that we have look into for make better usage of this product.
- Which company has better reach in market and amount of satisfaction user gets.
- On which products our company should concentrate more to have better reach.

Future Improvements

- We can make a website to have better vision and access for comparison of products.
- We can analyse comments much better to know at which point our company product lags or at which point it stood out of box comparing to competitors.
- We can take input datasets from other websites too for more accuracy.
- We can compare with more than two competing companies like nutanix, oracle etc for a product.

Resource Page

Data - https://www.gartner.com/en

Data Scraper - Instant Data Scraper Tool

Data Cleaning - https://medium.com/@yogeshojha/data-preprocessing-75485c7188c4

Data Visualization- Matplotlib, Seaborn Documentation

Data Analysis- https://www.sas.com/en_in/insights/analytics/machinelearning.html#:~:text=Machine%20learning%20is%20a%20method,Importance

Packages Used

- NLTK, Spacy
- pandas , PyDrive
- TextBlob
- Matplotlib, Seaborn
- WordCloud

THANK YOU !