

| **COURSE NAME:** | **Introduction to Data Science** |  |
| --- | --- | --- |
| **COURSE CODE:** | **AACS1573** |  |
| **ASSIGNMENT TITLE:** | **Data Science Application** |  |
| **SEMESTER:** | **202301** |  |
| **SUBMISSION DEADLINE:** | **Week 14 Monday before 12pm** | |
| **TOTAL MARK AWARDED:** | **Task Report Total:** | **80%** |
|  | **Presentation Total:** | **20%** |
|  | **Assignment Total:** | **100%** |
| **WEIGHTAGE TO FINAL MARK:** | **60%** |  |
| **STUDENT NAME and Student ID** | 1. **Jerome Lu Zheng Yao (2203589)** 2. **Kwek Wan Ying (22WMD03155)** 3. **Ng Mei Teng (22WMD03284)** |  |

**AACS1573 Introduction to Data Science**

**LEARNING OUTCOMES**

**Upon completion, students are expected to achieve the followings: -**

| **CLO2:** | **Interpret data into actionable insights that are found through data science process (C3,**  **PLO5)** |
| --- | --- |
| **CLO3:** | **Perform Exploratory Data Analysis (EDA) to analyse the main characteristics of datasets. (P3, PLO3)** |

**GUIDELINES TO STUDENTS**

**Students are required to observe the followings: -**

1. Students will form into groups of **3 - 4** members per group. Every team member is expected to contribute and participate actively in the entire process of completing the assignment. Sharing of ideas and assistance in the completion of assignments among members is required.
2. Upon completion of each assignment task, students are required to prepare an assignment task report.
3. All references and citations shall use the Harvard Referencing Style.
4. The assignment should be typed using 1.5 spaces between lines in 12 Times New Roman point- font, not more than 5 pages.
5. Plagiarism is strictly prohibited. Marks are awarded for your own (original) analysis. Therefore, use the time and information to build a well-constructed report.
6. Check carefully the submission date and the instructions given with the assignment. For late submission, there will be a reduction of absolute marks from the mark’s score submitted:

i. Late 1 to 3 days after the deadline of submission: minus 10 marks;

ii. Late 4 to 7 days after the deadline of submission: minus 20 marks;

iii. Late more than 7 days after the deadline of submission: 0 marks

1. The presentation will be on Week 14 (Duration at least 5 to 10 minutes per student depending on the lecturer’s allocation).
2. Presentation requirement including communication skills, visual aids and personal grooming are the criteria for obtaining the marks allocated.
3. Assignment format:
   1. Front Cover
   2. Assignment Marking Criteria
   3. Table of content
   4. Content following the Name of the assigned tasks and the corresponding task number.
   5. References and Appendix after each assignment task report.

**ASSIGNMENT**

**Introduction**

Data Science has a wide variety of applications. It is used in several fields ranging from health, education to transportation, and manufacturing. Various industries are using Data Science to boost their production, make smarter decisions, and develop innovative products that are tailored for customer needs.

Example of data science case studies from the following fields but not limited to:

* Manufacturing
* Pharmaceutical Industries
* Biotech
* Education
* Business

**Task**

* You are required to choose one of the above data science case study fields and describe clearly the data science application you have chosen.
* Your documentation should consist of the following:
  1. Introduction (include definition and significance of data science)
  2. Case study: Name and description (field and application)
  3. Data Science Process for the selected case study
  4. Results and Discussion
  5. Conclusion with the advantages and disadvantages of the selected case study • Present your assignment during week 12 (practical class).

**END OF ASSIGNMENT**

**MARKING CRITERIA**

Your written assignment and presentation will be assessed against the following criteria.

**ASSIGNMENT RUBRICS (CLO3)**

| **No.** | **Criteria** |  |  | **Evaluation/Marks** | |  |  | **Mark Achieved** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0 – Nothing**  **Presented/**  **Plagiarized** | **2 – Developing** | **4 – Approaching Expectation** | **6– Complete** | **8 – Excellent** | **10 – Beyond**  **Expectation** |
| 1. | **Logic and**  **Organisation** | Does not  develop ideas cogently,  uneven and  ineffective overall organisation, unclear introduction and conclusion | Still developing ideas cogently, uneven and ineffective overall organisation, unclear introduction and  conclusion | Develops unified and coherent ideas within paragraphs  with generally adequate transitions; clear  overall organisation relating most ideas together with good introduction and conclusion | Develops ideas cogently organises them logically with paragraphs and connects them with effective transitions Clear and specific introduction and  conclusion | Strongly Developed ideas cogently organise them logically with paragraphs and connect them with effective transitions Clear and specific introduction and  conclusion n | High innovative idea cogently organised and  logically sequenced with a clear and specific conclusion. |  |
| 2. | **Conceptual**  **Understanding** | Does not respond using course content | Respond using appropriate and sufficient course content | Responds using  appropriate and sufficient course  content | Respond clearly and effectively using appropriate and sufficient course content and outside sources | Good Response in using course content and resources in explaining your  understanding of the topic | Highly effective response in using course content and sources to  explain the  understanding |  |
| 3. | **Evidence** | Does not  present data | Presents accurately some of the  necessary data | Presents clearly and accurately most of the  necessary data | Presents clearly and accurately all of the necessary data | Good Sequence and  Logical thinking behind  Presenting data | Most suitable content explaining the end to end points and with clarity |  |
| 4. | **Relevance of Content** | Irrelevant content | Partially relevant content is  mentioned | Good relevant content is  mentioned | Right appropriate  Content is mentioned | Right content with clear logic content is mentioned | Highly effective content explaining the clear logic behind the topic |  |

| **No.** | **Criteria** | **Evaluation/Marks** | | | | | | **Mark Achieved** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0 – Nothing**  **Presented/**  **Plagiarized** | **2 – Developing** | **4 – Approaching Expectation** | **6– Complete** | **8 – Excellent** | **10 – Beyond**  **Expectation** |
| 5. | **In-text citation and Referencing** | No in-text citation and no  References | There are in-text citation or  References | There are in-text citations and  References but the  citation and references were not matching.  References were not using the  appropriate format. | There are in-text citations and References. The citation and reference were tallying. References were not using the appropriate format. | There are in-text citations and References. The citation and reference were tallying.  References were using an appropriate format. | There are in-text  citations,  References,  Bibliography. The  citation and reference were tallying.  References were using the  appropriate format. |  |
| 6. | **Use of Language** | Imprecise or inappropriate  choice of  words | Express thoughts  marginally | Appropriate choice of words | Uses rich choice of  words and  imaginative language | Good mix of words expressing the logic of thoughts clearly | Strong Effective use of Language expressing ideas  clearly |  |
| 7. | **Conclusion** | Does not draw conclusion or inference | Draws valid conclusions or  inferences | Draw valid  conclusions or  inferences  supported by  content | Draws clear and valid  conclusions or inferences supported by content | Good inferences by relevant content representation and inference | Effective conclusion based on right content representation and inference |  |
| 8. | **Reading/**  **Research** | No research has been done | Non-relevant research | Relevant research | Relevant research displayed in the content | Relevant research displayed in the content with appropriate examples and evidence | Relevant research displayed in the  content with  appropriate  examples and evidence and recommendation. |  |
| **TOTAL MARKS (80%)** | | | | | | | |  |

**PRESENTATION RUBRICS (CLO2)**

| **No** | **Evaluation**  **Categories** | **0 - No submission** | **1 – Weak**  **/Unsatisfactory** | **2 – Developing**  **/Needs Improvement** | **3 - Satisfactory** | **4 - Very Good** | **5 – Exceptional**  **/Excellent** | **Mark**  **Achieved** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **1.** | **INTRODUCTION** How well did the speaker set the scene | Did not appear/attend the presentation | Lack of interest and enthusiasm.  The presented topic was not introduced and explained.  Individuals were not introduced. | Shows a bit of interest and enthusiasm.  The presented topic was introduced but vague and irrelevant. Introducing self but not the members. | Seems interested, but could be more informed on the introduction.  All members were addressed, including self. | Interested and enthusiastic about the presentation. The introduction has been laid out properly and to the point.  All members were addressed, including self. | Eager about the presentation. Exceptional introduction which includes the summary.  All members were addressed, including self. |  |
| **2.** | **CONTENTS**  Was the objective identified? Was the presentation adapted to a wide range of audiences?  Was it well organized? | No clear statement offered.  Scope too broad or too narrow; lacks depth; AND uses too much technical language/ jargon.  No clear  information sequence; very difficult to follow. | Incomplete or unfocused.  Scope too broad or too narrow OR lacks depth OR uses too much technical language/ jargon. Evidence of some organization but not in an optimal order; difficult to follow. | Reasonably clear. Reasonable scope and depth; lapse into detail that may not be accessible to the audience.  Ideas presented in logical sequence; reasonably easy to follow. | Clear and concise.  Good scope & depth without  losing the audience in technical detail; a good learning experience.  Presented in logical & interesting ways; easy to follow but not oversimplified. | Clear, concise. Engaging, and thought-provoking. Exceptional scope & depth; a true learning experience; exceeds expectations. Exceptional organization because the topic is complex. |  |

| **No** | **Evaluation**  **Categories** | **0 - No submission** | **1 – Weak**  **/Unsatisfactory** | **2 – Developing**  **/Needs Improvement** | **3 - Satisfactory** | **4 - Very Good** | **5 – Exceptional**  **/Excellent** | **Mark**  **Achieved** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **3.** | **EVIDENCE**  Did the speaker demonstrate that actual work was carried out independently?  Was careful thought put into the work |  | No appropriate evidence was presented to support the  presentations central claims.  None of the  descriptions  contains an explanation.  Does not display a clear grasp of the subject being discussed. | Some evidence is present, but is either insufficient or not supportive of the main claims. Somewhat display some explanations but not clearly explain the significance of the central claims. | Evidence used to support the central claims is well chosen with some degree of detail.  Some of the descriptions contain explanations that clearly explain the significance of the central claims.  Displays a reasonable grasp of the subject. The opinion was clearly expressed. | Evidence well- chosen & detailed; the connection between argument & evidence is clear; opposing evidence considered. | Well-chosen, detailed, rich; highly compelling; opposing evidence considered and refuted.  Each description contains an explanation that clearly explains the significance of the central claims. Displays excellent understanding of the subject. Opinion  strongly expressed and supported. |  |
| **4.** | **CONCLUSIONS**  **CONFIDENCE**  **(reading materials)** How well did the author conclude, summarize and recommend? | No apparent conclusions; no discussion of implications. | Conclusions are restatements of previous statements. | Brings closure with some synthesis but does not address implications. | Synthesizes the work; brings  closure; allude to broader implications. | Synthesizes; brings closure; conveys real implications; suggest new perspectives. |  |
|  |  |  |  |  |  |  | **TOTAL MARKS (20%)** |  |

**Table of Content**

[**1. Introduction 9**](#_dkqxkhjlpmo3)

[1.1. Definition of Data Science 9](#_ympgd9bx23is)

[1.2. Significance of Data Science 9](#_iyy1948k4jsy)

[**2. Case Study 10**](#_gy7rzvg5881p)

[**3. Data Science Process 10**](#_cj5j85sq15i1)

[3.1. Data Preparation 10](#_orubo4ldxtby)

[3.2. Data Exploration 11](#_f2h0ajwr12g4)

[3.3. Data Representation 11](#_j889ygu9sugc)

[3.4. Data Discovery 11](#_xyskcpnosdx4)

[3.5. Learning from Data 12](#_b5r4yoazg0g8)

[**4. Result and Discussion 12**](#_cdg8whcnqxch)

[4.1. Category 12](#_jt1o8kvcxlbp)

[4.2. Size 13](#_6qpqwgtwv5o)

[4.3. Type 13](#_5j2od07qqoki)

[4.4. Price 13](#_bga5twn1c18r)

[4.5. Last Updated 14](#_qvnmsesm8e05)

[**5. Conclusions 14**](#_mn9zwmjc3dov)

[**Reference 16**](#)

[**Appendix 17**](#)

# Introduction

## Definition of Data Science

Data science is an interdisciplinary field that deals with data extraction, analysis, interpretation and visualization. It extracts insights and solves complex problems from data using technique and tools from statistics, mathematics, computer science and domain-specific knowledge. Data science is a broad field that includes everything from data cleaning and preparation to modeling and predictive analysis. It can be used to make informed decisions and gain a competitive advantage in a variety of industries, including business, healthcare, finance and government.

## Significance of Data Science

The significance of Data Science is to find the patterns within data. It is the fuel of the 21st century because it can apply to any field. Such as marketing, customer acquisition, innovation and enriching lives. For example, data science can be use in the business field. For business field, data science can let the manager level to do the smarter decisions and let them predict analysis to predict outcome.

In the discipline of data science, data is extracted, analyzed and interpreted to aid in decision making. Data science has become extremely relevant in today’s data driven world across a range of sectors and disciplines. Here are some of the significant benefits of data science. First and foremost, organizations may use data science to examine data to get insight and make wise decisions. Data science may give firms useful insight into consumer behavior, market trends, and other key data points that can aid in decision making. Beside that, increased productivity and efficiency are possible because of data science’s ability to automate a variety of jobs and procedures. For instance, using machine learning algorithms, operations like fraud detection, risk assessment and client segmentation may be automated as data is being analyzed.

Moreover, data science may assist businesses in tailoring their goods and services to the preferences of their clients. Organizations may comprehend the tastes, actions and demands of their consumers and adjust their products by evaluating customer data. Furthermore, organizations may use data science to areas where they can save expenses, streamline procedures and boost productivity. Over time, this may result in considerable cost reductions. Next, data science may assist businesses in locating new possibilities, developing fresh goods and services and creating fresh business strategies. Organizations may innovate and keep one stop ahead of the competition by using data science. In conclusion, data science is a vital discipline that significantly affects many different fields and industries. Organizations may use data science to make smarter decisions, increase productivity and efficiency, customize their services, save costs and innovate.

# Case Study

Name : Installations Of An Application In Google Play Store

Field : Application Development

Application : Knowing the total number of application installations helps teams measure success. Their marketing and growth methods are paying off if they are getting more installs, which signifies they are gaining more users. The objective of this analysis is to determine what is affecting the number of application installations.

In this case study, variables such as category, size, type, price, last updated may be the causes that affect the number of application installations.

# Data Science Process

## Data Preparation

Data preparation is the first step of the data science process that involves reading the data, data cleaning, data normalization and so on to prepare it for analysis. In this stage, we read the data to understand and to analyze the data we have chosen, which is 'Google Play Store Apps'. The fields contained in the dataset are category, size, type, etc. This process helps us to analyze the causes that affect the number of application installations and determine which data is useless.

Furthermore, cleansing the data is a very time-consuming part of data preparation (Zacharias Voulgaris, no date). To clean the data, we remove all the rows which have missing data since it only involves 13 rows (0.12%). Second, we drop all the duplicate rows. Third, in the Installs field, we remove the plus sign and comma to convert the data from text type into integer type. Fourth, in the Size field, we replace the metric symbol with its corresponding value. Fifth, in the Price field, we remove the dollar sign to convert the data from text type into integer type.

## Data Exploration

The purpose of data exploration is to gain a deeper understanding of the data, and to figure out the potential information that could be hiding within it (Zacharias Voulgaris, no date). By deeply exploring data, we will gain insight into the relationships between different variables, and have more effective decision making. In this stage, we can better understand which factors will affect the number of application installations through deeper analysis of the category, size, etc through data.

## Data Representation

Data representation basically involves assigning specific data structures to the variables involved (Zacharias Voulgaris, no date). In this stage, in the Type field, we represent the data with 0 and 1 since the variable only takes two values, which are Free and Paid. We represent ‘Free’ with 1 and ‘Paid’ with 0. Second, in the Last Updated field, we convert the data from string type into datetime format. After that, we represent the data with how long ago the application got updated in the last time (how many days ago) by finding the difference between the date of the application getting updated and the latest update of application in the dataset (8-8-2018).

## Data Discovery

Data discovery is the core of the data science process. Data discovery involves finding patterns in a dataset through hypothesis formulation and testing.(Zacharias Voulgaris, no date) In this stage, we make use of several statistical methods to prove the significance of the relationships that we observe. As what we observed and judged, category, size, type, price and last updated have a relationship with the number of application installations. Also, we apply some data visualization techniques such as pie charts and regression plots to discover data relationships that are otherwise not easily observable by looking at the raw data.

## Learning from Data

Learning from data is a crucial stage in the data science process and involves a lot of intelligent and creative analysis of a dataset using statistical methods and machine learning systems (Zacharias Voulgaris, no date). The results of learning from data can be used to make predictions on new data or to inform decision making. For example, based on our topic, the results of learning from data can be used to determine user behavior, identify new user segments, or to identify what aspects of the application should focus on or improve to attract more users.

# Result and Discussion

## Category

The pie chart (Appendix 2a) shows the categories of the applications in the store. There are 33 application categories. The top 5 app categories in the Google Play Store are Family (18.8%), Game (10.8%), Tools (8.1%), Business (4.1%) and Medical (3.9%). Other categories accounted for 54.2%. Therefore, the Family category has the most appearances for applications in store.

Based on the relationship shown (Appendix 2b), the Entertainment category has the highest average number of application installations, followed by Game and Photography. It showed that applications in the Entertainment category are downloaded by the most people on average.

## Size

The size field description and size distribution graph (Appendix 3a) show the distribution of the size field. The smallest size of the application is 8.5×103B/8.5KB and the largest size of the application is 1×108B/100MB. Average size is 2.13×107B/21.3MB. According to our observation, the size of most of the applications is around 1×107B/10MB.

Based on the regression plot shown (Appendix 3b), the larger the size, the higher the number of application installations. It seems like people prefer larger size applications. This may be because larger size applications have more and better features and functionality.

## Type

The pie chart of type distribution (Appendix 4a) shows that applications are separated into two groups based on their type, which are Free and Paid. Most of the applications in this store are free (93.1%) while 7.4% of the applications are paid.

The regression plot (Appendix 4b) shows the relationship between type and installs. In the type field, 0 indicates Paid and 1 indicates Free. Based on the graph shown, the number of application installations is higher if the application is free.

## Price

The price field description and price distribution graph (Appendix 5a) show the distribution of the price field. The average price is around 1.03 dollars, because most of them are free. The most expensive application is 400 dollars.

Based on the relationship shown (Appendix 5b), the higher the price, the lower the number of application installations. It seems like people are not very willing to spend a lot of money to purchase an application.

## Last Updated

From the last updated field description and last updated box plot (Appendix 6a) show the distribution of the last updated field. Among all the applications, the latest updated application was 0 days ago, and the oldest updated application was 3001 days ago, which is about 8 years ago. The average day an application was updated was 266 days ago, which is about 8 months ago. Furthermore, most of the applications were updated around 80 days ago.

Based on the regression plot shown (Appendix 6b), the latest application has a higher number of application installations. This may be because the more up-to-date application has met the latest requirements of the user.

# Conclusions

In conclusion, the factors that may influence the number of application installations are category, size, type, price and last updated days. There are some advantages and disadvantages of this case study.

The advantage of this case study is being able to provide static data to prove that those variable impacts may influence the number of application installations. Static data can help to make accurate calculations or reports based on category, size, type, price and last updated days which may influence the number of application installations.

The second advantage is that it can provide some references. For instance, what should be paid attention to, which aspects should be strengthened, and so on when developing a new application or improving the existing application to get more installations.

Nonetheless, there are some disadvantages. First, the data is from four or five years ago, which is different from the current situation, so it cannot fully represent the current situation. For example, the Entertainment category had the highest average number of installations five years ago, but it is not necessarily the same now.

Second, the number of application installations is not only affected by these variables, other reasons may have a greater impact on it, such as the function of the application, the user's favorability of the application, the degree of promotion of the application, and so on.

# Reference

1. LAVANYA, 2019, Google Play Store Apps, viewed on 13 March 2023, <https://www.kaggle.com/datasets/lava18/google-play-store-apps>
2. Zacharias Voulgaris, no date, Data Scientist: The Definitive Guide to Becoming a Data Scientist, viewed on 3 May 2023, https://books.google.com.my/books?id=yIvXBgAAQBAJ&printsec=frontcover#v=onepage&q&f=false

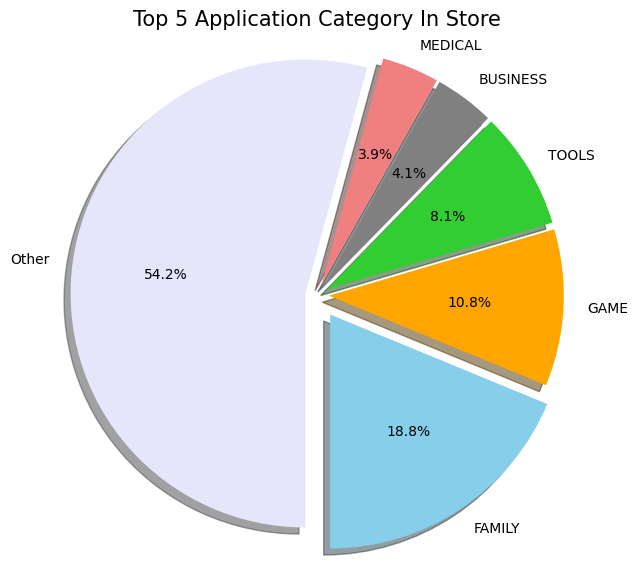
# Appendix

Appendix 1 - Installs Category

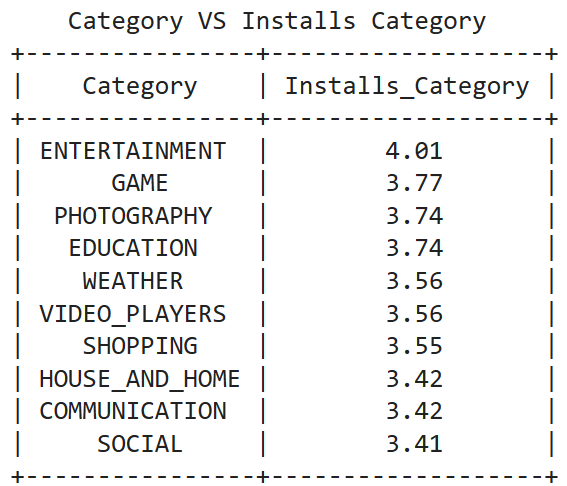
| **Installs Category** | | |
| --- | --- | --- |
| **No.** | **Represent** | **List of number of installations** |
| 1 | Very Low | 0, 1, 5, 10 |
| 2 | Low | 50, 100, 500, 1000 |
| 3 | Medium | 5000, 10000, 50000, 100000 |
| 4 | High | 500000, 1000000, 5000000, 10000000 |
| 5 | Very High | 50000000, 100000000, 500000000, 1000000000 |

Appendix 2 - Category

1. Top 5 Out Of 33 Application Categories

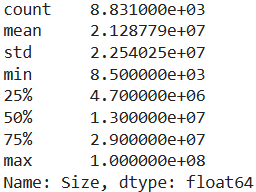


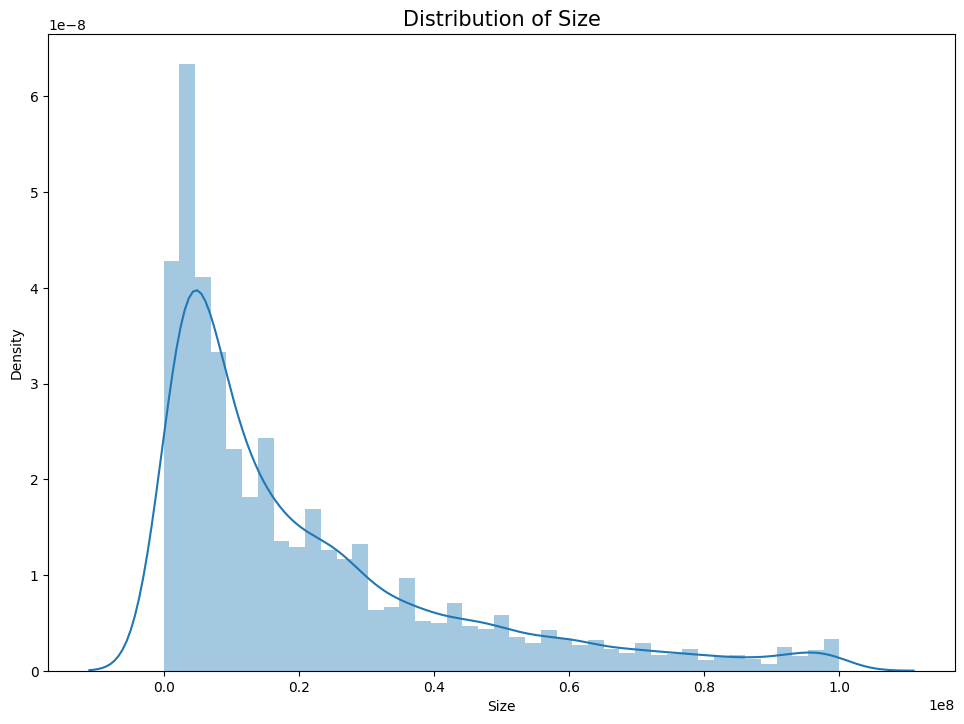
1. Relationship Between Category And Install\_Category



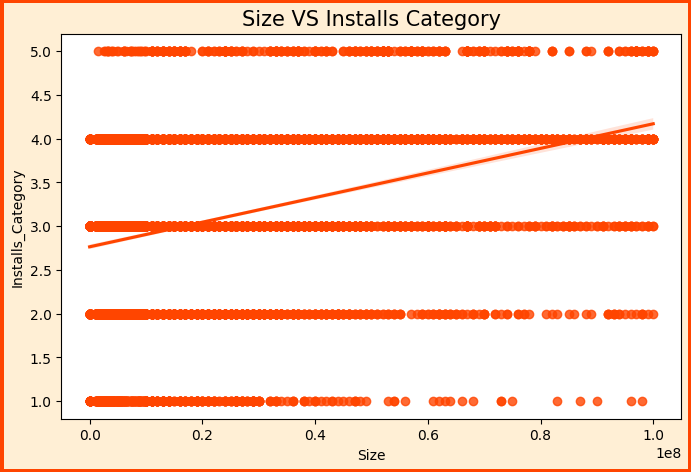
Appendix 3 - Size

1. Size Distribution



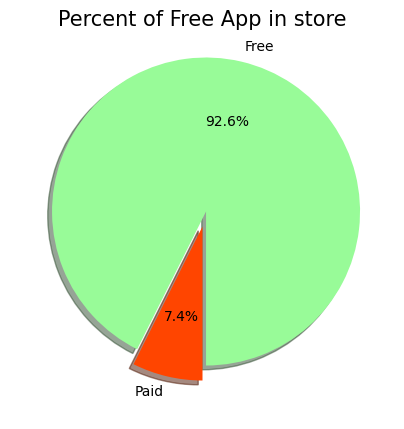


1. Relationship Between Size And Install\_Category

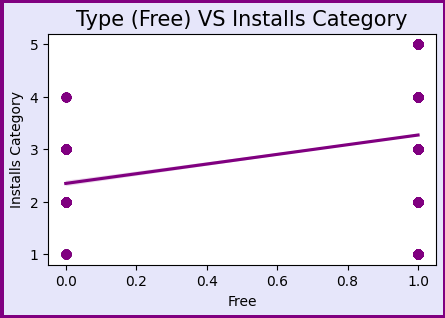


Appendix 4 - Type

1. Percent Of Free App In Store

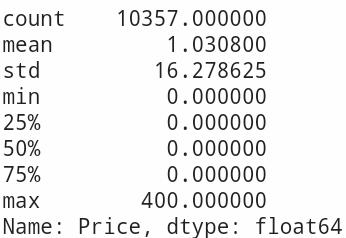


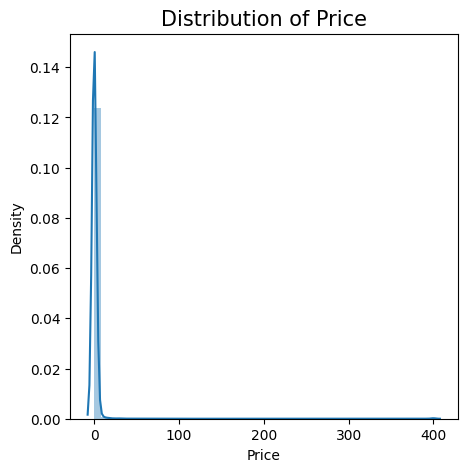
1. Relationship Between Type And Install\_Category



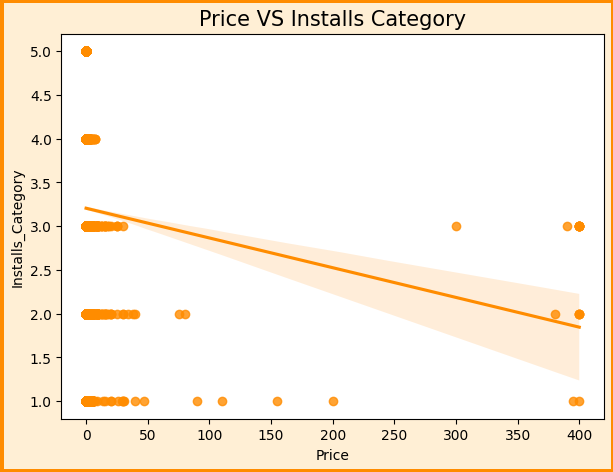
Appendix 5 - Price

1. Price Distribution



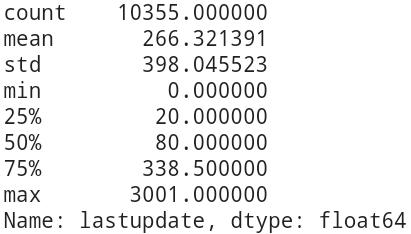


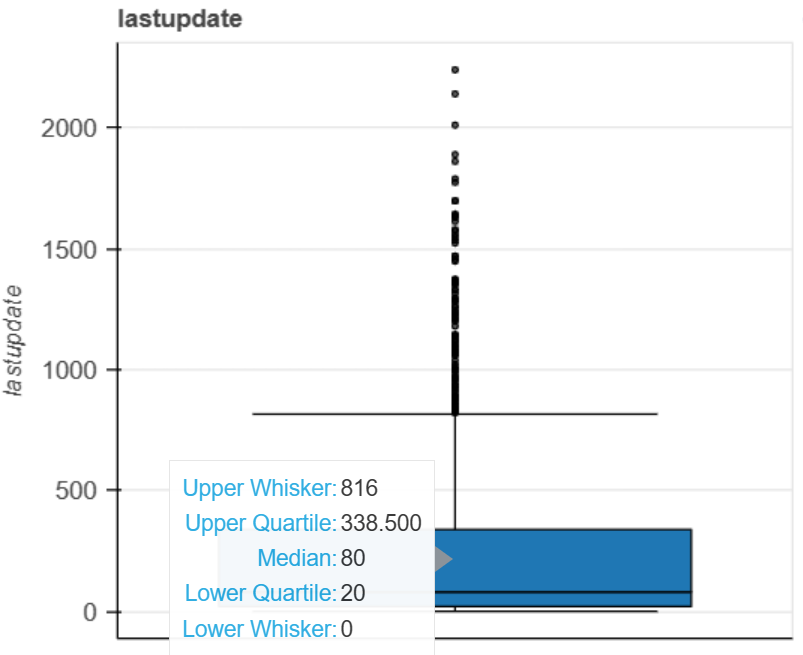
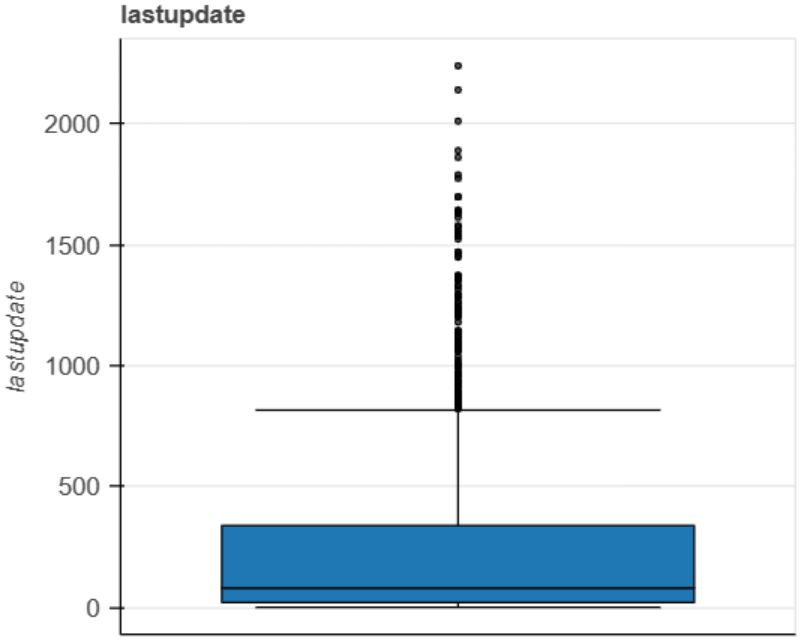
1. Relationship Between Price And Install\_Category



Appendix 6 - Last Updated (days ago)

1. Last Updated (days ago) Distribution





1. Relationship Between Last Update (days ago) And Install\_Category

