WM0824 - Economics of Security

# Relation Between Security Level and Application Market Categories

November 14, 2016

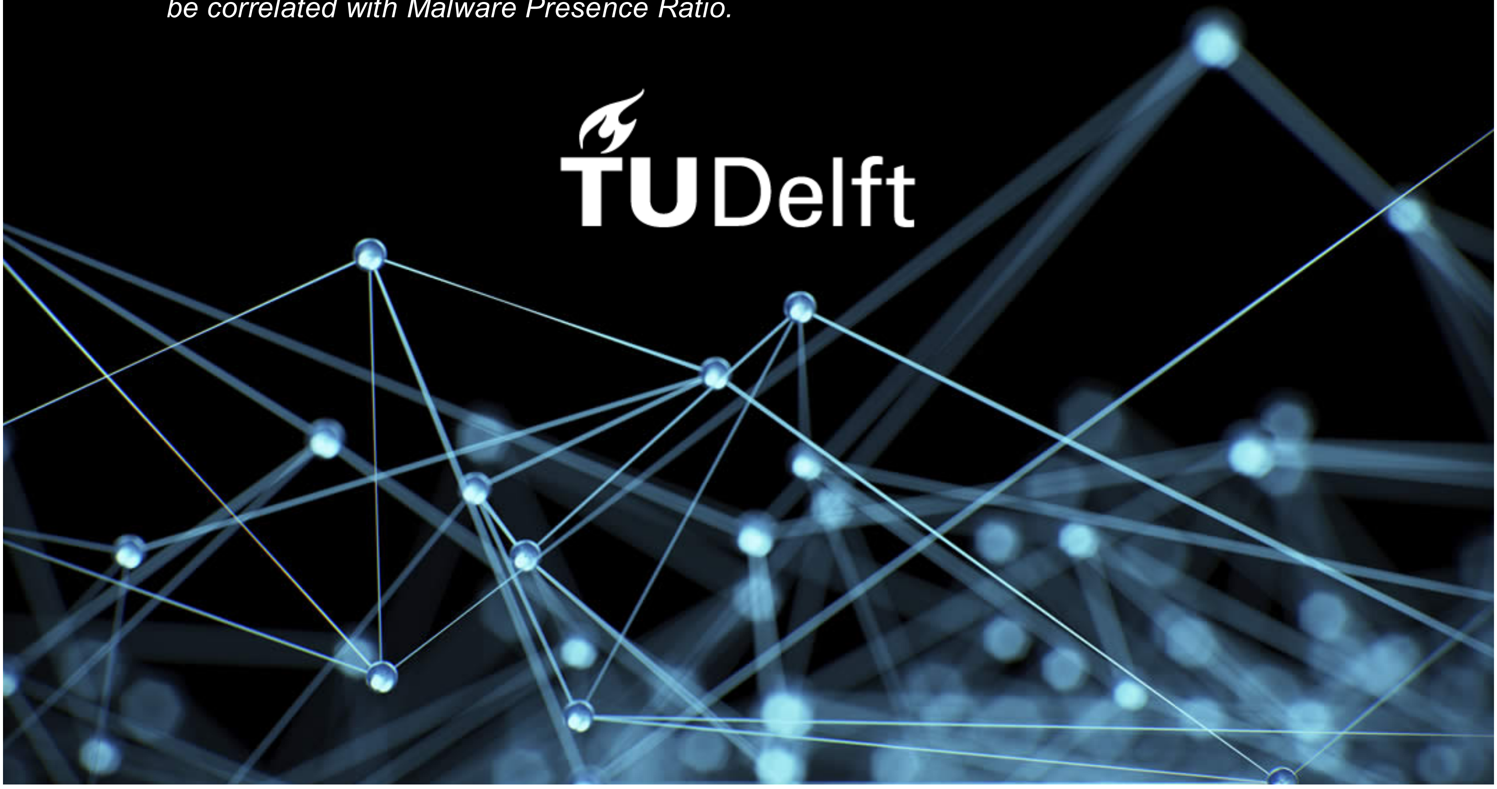
Raditya Arief

4500318

## Abstract

*In this report, the author attempted to gain further understanding about security level in various categories of apps across Android app marketplace. To guide this research, a research question is developed: “Are there any factor that influence the difference of security level in various categories of app market?” The two formulated hypotheses are (1) “There are significant differences on security level difference between categories in the app market” and “There is a significant correlation between the number of downloads of a particular app category and its security level metrics.”*

*The author has found that there are significant differences between means of security level in each application category. The author has also found that the number of total download in each category correlated with Malware Download Ratio, one of the security level metrics mentioned in this report. The total download in each category, however, does not found to be correlated with Malware Presence Ratio.*



WM0824 - Economics of Security

# Relation Between Security Level and Application Market Categories

November 14, 2016

Raditya Arief

4500318

## Abstract

*In this report, the author attempted to gain further understanding about security level in various categories of apps across Android app marketplace. To guide this research, a research question is developed: "Are there any factor that influence the difference of security level in various categories of app market?" The two formulated hypotheses are (1) "There are significant differences on security level difference between categories in the app market" and "There is a significant correlation between the number of downloads of a particular app category and its security level metrics."*

*The author has found that there are significant differences between means of security level in each application category. The author has also found that the number of total download in each category correlated with Malware Download Ratio, one of the security level metrics mentioned in this report. The total download in each category, however, does not found to be correlated with Malware Presence Ratio.*

# 1. Introduction

Android is currently one of the most popular mobile platforms in the world. Their popularity can easily be justified in term of download numbers. The total number of downloads for Android platform from the Google Play store and its third-party app stores looks to dominate over the number of downloads for Apple iOS as presented in Figure 1. However, a large part of this download volume has to be accredited to few Chinese-based app markets. These app markets are not only a majority in Android app download volume share but they are also make up a large part of the global app market. In fact, these three markets are so big that they accounted for more than 50% of global mobile app downloads in 2014 (Perez, 2015). This statistic reveals just how significant these Chinese-based app markets are.

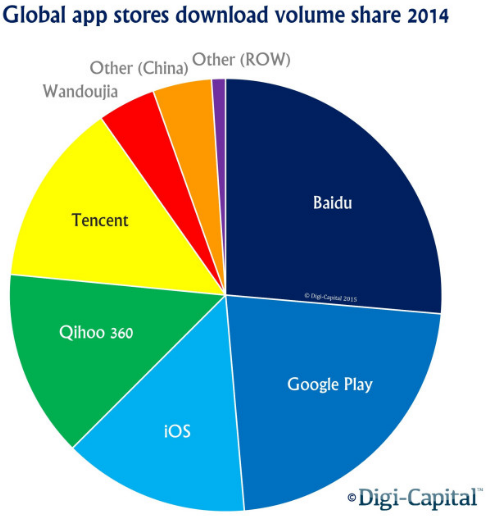


Figure 1. Global app stores download in 2014

However, the third-party Android app markets are also well-known for its security issues. One research tried to measure the ratio of malware infected apps in the Google Play store as well as several third-party app markets. (Kikuchi, Mori, Nakano, Yoshioka, Matsumoto, & Van Eeten, Evaluating Malware Mitigation by Android Market Operators, 2016). According to the research, the Google Play store has a malware infection ratio of 0.1%, while the infection rate of the rest of the third-party app markets are ten times higher. Considering the severity of the malware infection problem coupled with the size of the market, the impact of this security issue must have been quite substantial. Therefore, any effort in the direction of mitigating this issue should be considered highly important.

### The Preceding Reports

This report is a continuation of a series of assignments, made by Group 5 of the Economics of Security class, which discussed the problem of malware infection in Chinese third-party Android markets. The previous assignments have discussed different problems related to this topic. The Android market operator was defined as the main problem owner. The first assignment described the underlying security issue of the datasets and security metrics that can be developed. The second assignment discussed the risk strategies that can be taken by various actors to mitigate the issue.

The third assignment described the countermeasure that could be taken by the actors and whether the actors have an incentive to pursue the countermeasure. The third assignment also discussed the underlying factors that might have affected the security performance.

The analysis of these assignments were based on four datasets of applications drawn from two different third-party Android app markets: Baidu and Qihoo 360. The applications in the datasets are samples taken from various categories in these Chinese-based markets. The structure of the datasets is described in Table 1.

**Table 1**. The structure of Android app repository datasets

| **Column** | **Description** |
|---|---|
| Market | The market from which the app was hosted (Baidu or Qihoo 360) |
| Category | The market category for the app |
| AppName | The name of the app |
| Version | The release version of the app |
| Size | The size of the installer package |
| UpdateDate | The date of the latest update of the app |
| Package | The name of the installer package |
| DownloadTImes | The number of how many times the app has been download |
| DownloadUrl | The download link for the app in the market |
| DownloadResult | A Boolean value of whether the download was a success or failure |
| DownloadTime | The time and date when the app was accessed/downloaded |

As can be seen in Table 1, initially the datasets did not contain direct information about any security issue, so the team had to devise a procedure to produce an insight regarding the security problem. For this series of assignments, malware infection was chosen to be the main security issue. Three different security metrics were developed to represent the security level against malware infection. The first metric, the Malware Detection Ratio, represents the average of malware infections in every app in a particular market. The second metric, the Malware Presence Ratio, is the rate of malware-infected apps in a particular market. The last metric, the Malware Download Ratio, measures the ratio of malicious downloads divided by the total download in a particular market. These metrics will be used for the analysis in this report. From this point onward, these metrics will be referred simply as security level.

### Purpose of This Report

This report will try to extend this research series a little bit further. As mentioned before, the third assignment have discussed the factors that might affect the security performance of a particular market, which reflected on the difference of the security metric. Interestingly, a brief look at the datasets reveals meaningful differences of the aforementioned security metrics between various categories of the app markets. This finding requires a further research before any meaningful conclusion can be deduced.

Thus, it becomes the purpose of this paper to gain further understanding in regard to the security metric differences among various app market categories. Understanding this differences might yield insights that can support the effort to mitigate the security risk in discussion.

For example, in the previous assignment it has been found that the distribution of cost and benefits of secure software development is unbalanced for the app developers. The network effect characteristic in the world of information technology has forced the developers to abandon costly secure software development in order to pursue market dominance. Obviously, the problem with security will affect the users of the app. This imbalance presents a significant externality problem.

Thus, understanding these factors might help with the allocation of costs and benefits among different types of app developers. Identifying the factors that affect the security level differences among the categories and how these factors affect them will be the main goal of researches in this direction. However, this goal will not be in the scope of this report because of various limitations that will be discussed later.

## 2. Literature Review

There are five literatures that found to be strongly relevant with the security issue under investigation.

**Detecting Repackaged Smartphone Applications in Third-Party Android Marketplaces**

Zhou, Zhou, Jiang, and Ning (2012) examined the presence of repackaged apps in six third-party android application markets. Repackaged apps are those apps from the Android Market (now Google Play store) that are repackaged and re-distributed to the third-party marketplaces for whatever reason. To measure these repackaged apps, the authors presented a system called DroidMOSS, which works by measuring the similarity between two apps. To do the analysis, they took 200 apps from each of the third-party marketplaces and measure those apps against 68,187 apps taken from the official Android Market. They found that between 5% to 13% apps from the third-party marketplace are repackaged. A further investigation reveals that these apps are mainly repackaged to replace existing in-app advertisement. There are also more serious findings where apps are planted with backdoors or malicious payloads.

**Evaluating Malware Mitigation by Android Market Operators**

Kikuchi, Mori, Nakano, Yoshioka, Matsumoto, and Van Eeten studied the malware mitigation effort in Google Play store and four third-party app markets. First, the authors define three security metrics to measure the mitigation effort. Next, samples are downloaded from each markets. Afterwards, the downloaded apps are submitted to VirusTotal and retrieve the results. It is found that Google Play had malware presence ratio of 0.1%, while the third-party app markets had ten times over higher than in the Google Play store. Google Play store is considered as the only app market with active malware removal.

**Dissecting Android Malware: Characterization and Evolution**

Zhou and Jiang (2012) aimed to systematically characterize Android malware based on their behavior. The authors gathered 1260 malware samples from 49 different malware families and studied different aspects of their behavior, such as installation method, activation mechanism, and its malicious payloads. There are several key findings in this journal First, the authors found that 86% of malware samples are legitimate apps that has been repackaged with malicious payload added into it. Second, around 36.7% of the malware samples utilizes root-level exploit. Third, 93%

of the malware samples have bot-like capability. Moreover, the authors also suggest that the solutions for these problems are severely lagging.

**Category Based Malware Detection for Android**
Grampurohit, Kumar, Rawat, & Rawat (2014) suggested that the accuracy of machine learning technique to detect malware can be improved by using market category information. This paper recognized 30 app categories on the Google Play store, such as Social, Communication, Education, etc. The authors found that the malware detection accuracy achieved an average improvement of 3 – 4% by using market category information.

**Android Malware Detection Using Category-Based Machine Learning Classifiers**
Alatwi (2016) also proposed category-based machine learning to increase the accuracy of malware detection. The author stated that applications under the same category tend to share a common set of features. Machine learning technique can differentiate the "abnormal" request that is uncommon for an app in a particular category, indicating the presence of malware. The result of category-based classifier showed a remarkable increase in performance compared to the non-category based.

The findings in the last two researches have implicitly stated that there is indeed a relationship between malware characteristics and the market category where the malware is hosted. These findings indicate that there also might be a relationship between the security level of a market category and another factor related to the same category.

## 3. Research Question, Objective, and Hypothesis

As has been stated above, the purpose of this paper is to gain further understanding in regard to the security metric differences among various app market categories. The direction of this work is was taken with the realization that currently there are very little to none research done to understand the relation between the characteristics of an app in a particular category to the security level of the category itself. In this preliminary study, the author will try to investigate whether a research in this direction is worthy to undertake.

To answer the question, the following research question is developed:
"*Are there any factor that influence the difference of security level in various categories of app market?*"

To answer the main research question, the following sub-questions are developed.

**First sub-question**: "*Are there any significant difference on security level between application market categories?*"

**Second sub-question**: "*Are there any significant relationship between factors and the security level of app categories?*"

Although the lack of similar studies and the lack of data consequently have limited the goal of this work, but this research could potentially open a whole new research area and possibly allow for further research to be done.

**Hypothesis**
Based on the objective and literature review that has been done, there are three hypotheses formulated for this report:

**First hypothesis**: There are significant differences on security level difference between categories in the app market.

**Second hypothesis**: There is a significant correlation between the number of downloads of a particular app category and its security level metrics.

## 4. Methodology

The study in this report utilize quantitative research evaluation to answer the research question. This study will make use of secondary data source taken from an earlier report.

The data source is application download details taken from two different third-party Android marketplaces. In the previous study, a stratified sampling method was employed; the list of applications download information were clustered based their category. However, there was still the fear of bias because the popular apps tend to cleaner than the unpopular ones. To reduce the bias, these clusters are divided into groups based on their download numbers and from each groups samples are taken. The result of the earlier study is the evaluation of malware presence in every samples of application.

In this report, the resulting analysis from the previous study is further grouped based on their categories. There are two app markets each of which has its categorization structure. In order to come up with a unified dataset, the categories will have to be grouped according to its similarities. The result of the combined categories is presented in Table 2.

**Table 2**. Category grouping between Baidu and Qihoo 360

| Group | Baidu Market Category | Qihoo 360 Software Category |
|---|---|---|
| 1 | Money Shopping | Shopping Offer |
| 2 | News Reading | News Reader |
| 3 | Office School | Office Business |
| 4 | Shooting Landscaping | Photography |
| 5 | Social Communication | Social Communication |
| 6 | System Tool | System Security |
| 7 | Video Player | Audio-Video |
| 8 | Theme Wallpaper | Theme Wallpaper |

Next, the data then will be run through analysis by utilizing SPSS Statistic software. There are two main analyses that will be done. First, one-way ANOVA will be used to test the difference means of security level between categories. Second, liner regression analysis will be used detect possible correlation between the total number of downloads of a particular category with its security level.

The lack of data is the main limitation in this study. It is highly possible that the analysis to find no correlation at all between the available factors because there are only few of these factors available for research.

# 5. Results

In this section, the author describes the results found during the analyses.

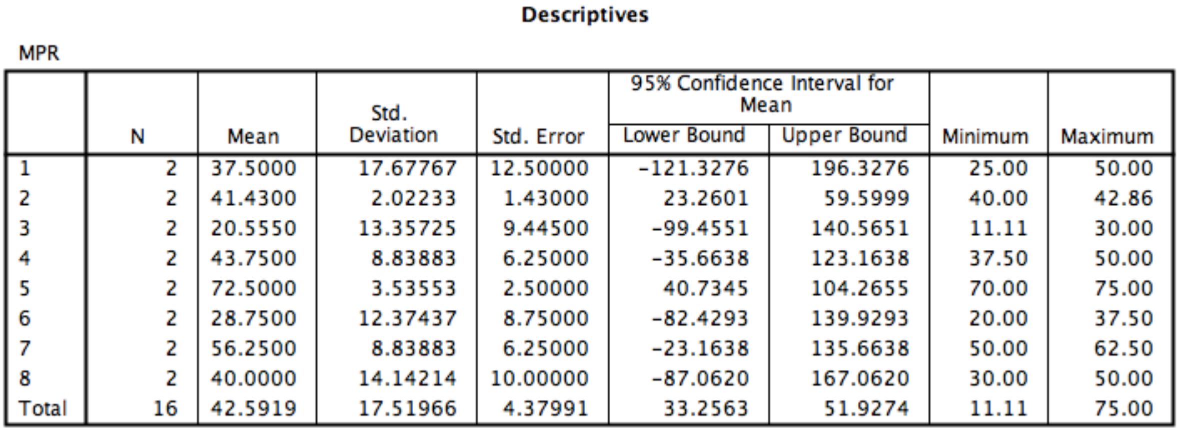
## 5.1. Security Level Difference Across Categories

The first test is to analyze the significance in difference of security level between various categories. As mentioned before, the term security level represents three security metrics that were developed in the previous study. For this study, only Malware Presence Ratio and Malware Download Ratio will be used to measure the security level. Malware Detection Ratio will not be measured because it is based on different unit of analysis; the Malware Presence Ratio and Malware Download Ratio measure the market while Malware Detection Ratio measure an app. Also already mentioned in the previous section, the categories from both Baidu and Qihoo 360 are combined into groups of categories. Having established these, the analysis can be run.

One-way ANOVA is used to test the significance of the security level between categories (or groups of categories). Both of the security metrics are used as the depended variable, while the groups of categories used as factor. For this study, statistical significance (or alpha) is set to 0.10, not the typical 0.05. The reason for choosing a larger alpha is because this test is intended to see the difference (in means) that ‘might’ exist. A smaller alpha of 0.05 or even 0.01 may be used in future study to more accurately detect the difference that ‘really’ exists.

Table 3 and Table 4 below present the result of one-way ANOVA on Malware Presence Ratio.

**Table 3**. Details on Malware Presence Ratio analysis

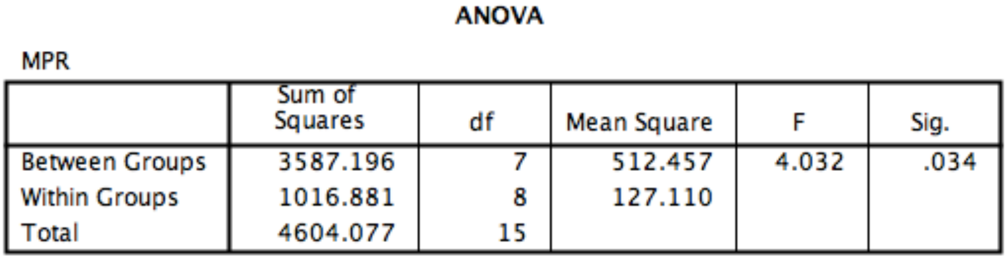


**Descriptives**

MPR

|  |  |  |  |  | 95% Confidence Interval for Mean |  |  |  |
|---|---|---|---|---|---|---|---|---|
|  | N | Mean | Std. Deviation | Std. Error | Lower Bound | Upper Bound | Minimum | Maximum |
| 1 | 2 | 37.5000 | 17.67767 | 12.50000 | -121.3276 | 196.3276 | 25.00 | 50.00 |
| 2 | 2 | 41.4300 | 2.02233 | 1.43000 | 23.2601 | 59.5999 | 40.00 | 42.86 |
| 3 | 2 | 20.5550 | 13.35725 | 9.44500 | -99.4551 | 140.5651 | 11.11 | 30.00 |
| 4 | 2 | 43.7500 | 8.83883 | 6.25000 | -35.6638 | 123.1638 | 37.50 | 50.00 |
| 5 | 2 | 72.5000 | 3.53553 | 2.50000 | 40.7345 | 104.2655 | 70.00 | 75.00 |
| 6 | 2 | 28.7500 | 12.37437 | 8.75000 | -82.4293 | 139.9293 | 20.00 | 37.50 |
| 7 | 2 | 56.2500 | 8.83883 | 6.25000 | -23.1638 | 135.6638 | 50.00 | 62.50 |
| 8 | 2 | 40.0000 | 14.14214 | 10.00000 | -87.0620 | 167.0620 | 30.00 | 50.00 |
| Total | 16 | 42.5919 | 17.51966 | 4.37991 | 33.2563 | 51.9274 | 11.11 | 75.00 |

**Table 4**. ANOVA analysis on Malware Presence Ratio



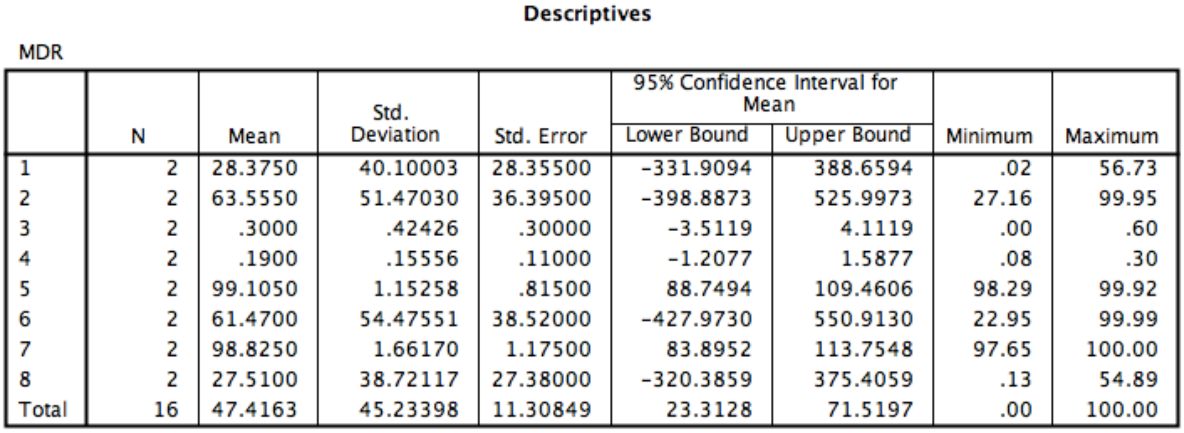
**ANOVA**

MPR

| | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Between Groups | 3587.196 | 7 | 512.457 | 4.032 | .034 |
| Within Groups | 1016.881 | 8 | 127.110 | | |
| Total | 4604.077 | 15 | | | |

The result from Malware Presence Ratio analysis is quite encouraging. In Table 4, it can be seen that the p-value (under the column Sig.) for this test is 0.034, well below the alpha of 0.10, which means the null hypothesis can be rejected. It can be concluded that the Malware Presence Ratio differences between groups of categories is statistically significant. The result of the analysis for Malware Download Ratio can be seen in Table 5 and Table 6.

**Table 5**. Details on Malware Download Ratio analysis

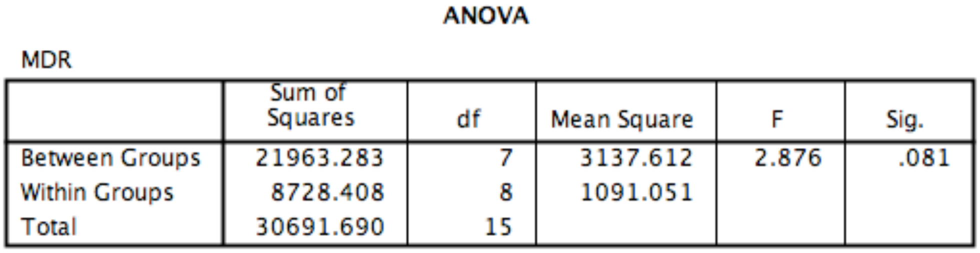


**Descriptives**

MDR

| | N | Mean | Std. Deviation | Std. Error | 95% Confidence Interval for Mean | | Minimum | Maximum |
|---|---|---|---|---|---|---|---|---|
| | | | | | Lower Bound | Upper Bound | | |
| 1 | 2 | 28.3750 | 40.10003 | 28.35500 | -331.9094 | 388.6594 | .02 | 56.73 |
| 2 | 2 | 63.5550 | 51.47030 | 36.39500 | -398.8873 | 525.9973 | 27.16 | 99.95 |
| 3 | 2 | .3000 | .42426 | .30000 | -3.5119 | 4.1119 | .00 | .60 |
| 4 | 2 | .1900 | .15556 | .11000 | -1.2077 | 1.5877 | .08 | .30 |
| 5 | 2 | 99.1050 | 1.15258 | .81500 | 88.7494 | 109.4606 | 98.29 | 99.92 |
| 6 | 2 | 61.4700 | 54.47551 | 38.52000 | -427.9730 | 550.9130 | 22.95 | 99.99 |
| 7 | 2 | 98.8250 | 1.66170 | 1.17500 | 83.8952 | 113.7548 | 97.65 | 100.00 |
| 8 | 2 | 27.5100 | 38.72117 | 27.38000 | -320.3859 | 375.4059 | .13 | 54.89 |
| Total | 16 | 47.4163 | 45.23398 | 11.30849 | 23.3128 | 71.5197 | .00 | 100.00 |

**Table 6**. ANOVA analysis on Malware Download Ratio



**ANOVA**

MDR

| | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Between Groups | 21963.283 | 7 | 3137.612 | 2.876 | .081 |
| Within Groups | 8728.408 | 8 | 1091.051 | | |
| Total | 30691.690 | 15 | | | |

The result from Malware Download Ratio is promising, although not as reassuring as the Malware Presence Ratio analysis. The p-value of the test is 0.081. This value is still below the set alpha of 0.10, which means the null hypothesis can be rejected.

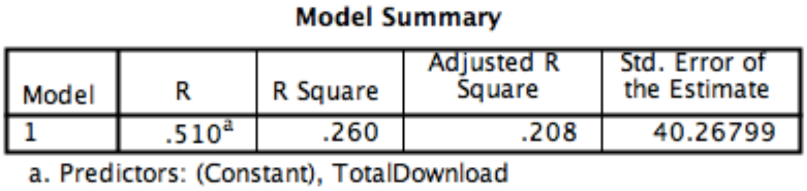
From these results, it can be inferred that there are significant security level differences between categories in both marketplaces. Moreover, this results also confirm the first hypothesis stated in Chapter 3.

## 5.2. Correlation Between Total Downloads and Security Level

In this second test, the author will try to test whether the result found in the previous test can be attributed to a particular factor, making this factor an independent variable for the dependent security metrics. In this case, the independent variable is the number of total downloads of each category. The number of total downloads is used because the lack of data seriously limits the available factors to choose from. In this test, the linear regression method will be used to test any correlation between the number of total downloads in each category and the security metrics.

Table 7 and Table 8 present the linear regression analysis for Malware Download Ratio.

**Table 7**. Summary of linear regression analysis between Total Downloads and Malware Download Ratio

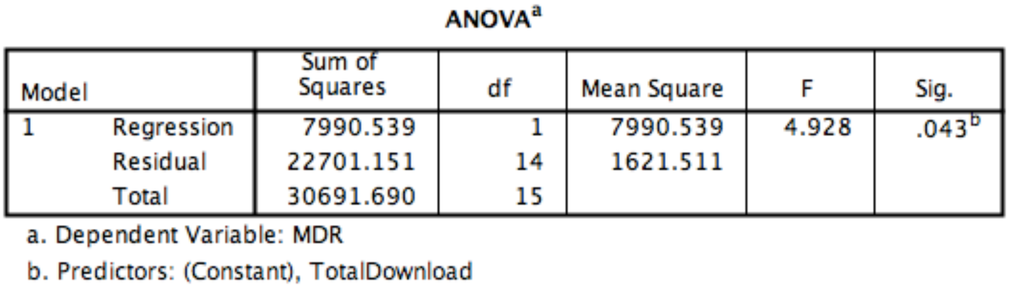


**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .510[a] | .260 | .208 | 40.26799 |

a. Predictors: (Constant), TotalDownload

**Table 8**. ANOVA table of linear regression analysis between Total Downloads and Malware Download Ratio



**ANOVA[a]**

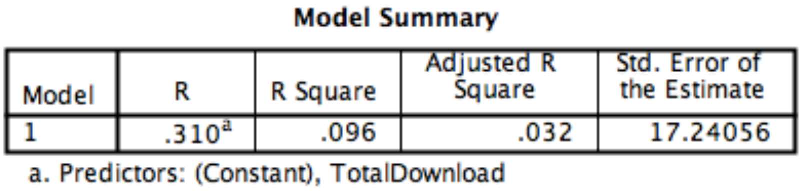
| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 7990.539 | 1 | 7990.539 | 4.928 | .043[b] |
| | Residual | 22701.151 | 14 | 1621.511 | | |
| | Total | 30691.690 | 15 | | | |

a. Dependent Variable: MDR

b. Predictors: (Constant), TotalDownload

The linear regression analysis for the number of total downloads and Malware Download ratio shows a surprising significant result. The p-value from the ANOVA table on Table 8 is 0.043, which means the null hypothesis of “no correlation between download number in categories with the Malware Download Ratio” can be rejected. The R Square from Table 7 indicates that 26% of total variance in Malware Download Ratio can be explained by the number of downloads in the category. This results support the second hypothesis.

**Table 9**. Summary of linear regression analysis between Total Downloads and Malware Presence Ratio

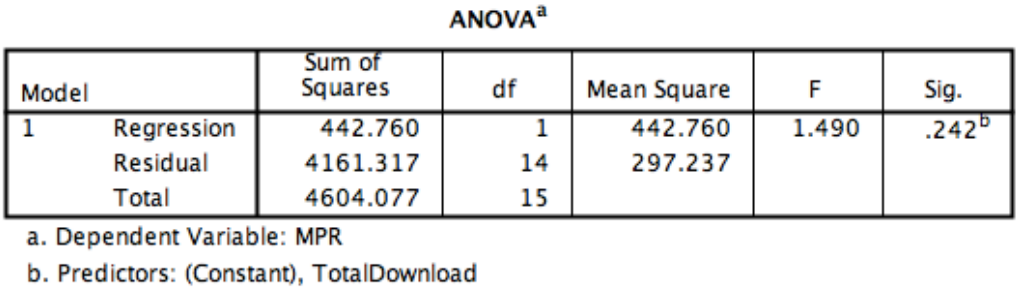


**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .310[a] | .096 | .032 | 17.24056 |

a. Predictors: (Constant), TotalDownload

**Table 10**. ANOVA table of linear regression analysis between Total Downloads and Malware Presence Ratio



**ANOVA**[a]

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 442.760 | 1 | 442.760 | 1.490 | .242[b] |
| | Residual | 4161.317 | 14 | 297.237 | | |
| | Total | 4604.077 | 15 | | | |

a. Dependent Variable: MPR

b. Predictors: (Constant), TotalDownload

Table 9 and Table 10 present the result of linear regression analysis of the number of downloads in categories and its Malware Presence Ratio. The p-value in Table 10 indicates insignificant result, therefore the null hypotheses of “there is no correlation between downloads and Malware Presence Ratio” cannot be rejected. Since there is no significant correlation, there is no use in interpreting Table 10.

In conclusion, although there is no significant correlation between the number of downloads with Malware Presence Ratio, a significant correlation with Malware Download Ratio is found. These findings support the second hypothesis. Moreover, these findings also answer the second sub- question of the main research question stated in Section 3.

## 6. Limitations

The author has done the study in this report with some important limitations in mind. First, the limitation of the previous study consequently influences this study. The utilization of VirusTotal in the first assignment implies that there is an unknown level of inaccuracy of malware infection rate in the data. VirusTotal has the tendency to keep an app flagged as malware even after an effort is made to remedy the infection. This inaccuracy consequently influences the Malware Presence Ratio value as well as the Malware Download Ratio value used in the analysis.

Another important limitation is the lack of data. The team who worked on the previous assignments received more than 50,000 rows of data about applications from the two markets. However, the VirusTotal Public API used to detect for malware infection is restricted to two inspection per-minute. This API limitation coupled with the limited time resulted in just over 400 rows of application data with its security examination. The lack of data can also be seen from the perspective of the number of markets. Since there are only two markets in the data, every statistical analysis done in this paper relies on the data of these markets. Two sample generally considered insufficient for many statistical analyses. More data from different app market would certainly be helpful.

The second limitation leads to another limitation is related to the statistical analysis method used in this study. Statistical method usually relies on some assumptions and the significant lack of data may prohibit the achievement of these assumptions. This limitation might have seriously affected the result of the analysis.

## 7. Conclusion

In this report, the author attempted to gain further understanding about security level in various categories of apps across Android app marketplace. The analysis in this report are based on the data of two different third-party Android app marketplaces. The different categorization employed by the two marketplaces have been combined into eight group of categories. There are two main findings in this report. First, the author has found that there are significant differences between means of security level in each application category. Second, the author has also found that the number of total download in each category correlated with Malware Download Ratio, one of the security level metrics mentioned in this report. The total download in each category, however, does not found to be correlated with Malware Presence Ratio.

The results obtained in this report indicates a possibility that there are indeed factors that influence the difference of security level across different categories in app marketplace. The presence of factors influencing the security level of certain part of app marketplace potentially change the way the problem owner mitigates the security issue. However, the author does recognize that some serious limitations may have reduced the validity of the result. Nevertheless, the author urges that further studies with more app marketplace samples and a more complete information regarding each marketplace to be done in order to confirm the indication found in this report.

# References

Woods, B. (2016, January 20). *Google Play had twice as many app downloads as Apple's App Store in 2015* . Retrieved November 6, 2016, from The Next Web: http://thenextweb.com/apps/2016/01/20/google-play-had-twice-as-many-app-downloads-as-apples-app-store-in-2015/

Kan, M. (2015, February 25). *With no Google, Chinese app stores soar on high downloads*. Retrieved November 6, 2016, from PC World: http://www.pcworld.com/article/2888892/with-no-google-chinese-app-stores-soar-on-high-downloads.html

Perez, S. (2015, April 27). *Android Surpasses iOS In Revenue, If China's Android App Stores Are Combined*. Retrieved November 6, 2016, from Tech Crunch: https://techcrunch.com/2015/04/27/android-surpasses-ios-in-revenue-if-chinas-android-app-stores-are-combined/

Kikuchi, Y., Mori, H., Nakano, H., Yoshioka, K., Matsumoto, T., & Van Eeten, M. (2016). Evaluating Malware Mitigation by Android Market Operators. *9th Workshop on Cyber Security Experimentation and Test (CSET 16)* (pp. 1-8). Austin, TX: USENIX Association.

Zhou, W., Zhou, Y., Jiang, X., & Ning, P. (2012, February). Detecting repackaged smartphone applications in third-party android marketplaces. *Proceedings of the second ACM conference on Data and Application Security and Privacy* , 317-326.

Zhou, Y., & Jiang, X. (2012, May). Dissecting android malware: Characterization and evolution. *IEEE Symposium on Security and Privacy* , 95-109.

Kikuchi, Y., Mori, H., Nakano, H., Yoshioka, K., Matsumoto, T., & Van Eeten, M. (n.d.). Evaluating Malware Mitigation by Android Market Operators. *9th Workshop on Cyber Security Experimentation and Test (CSET 16)* .

Grampurohit, V., Kumar, V., Rawat, S., & Rawat, S. (2014, September). Category Based Malware detection for Android. *International Symposium on Security in Computing and Communication* , 239-249.

Alatwi, H. (2016). *Android Malware Detection Using Category-Based Machine Learning Classifiers.* Rochester Institute of Technology.