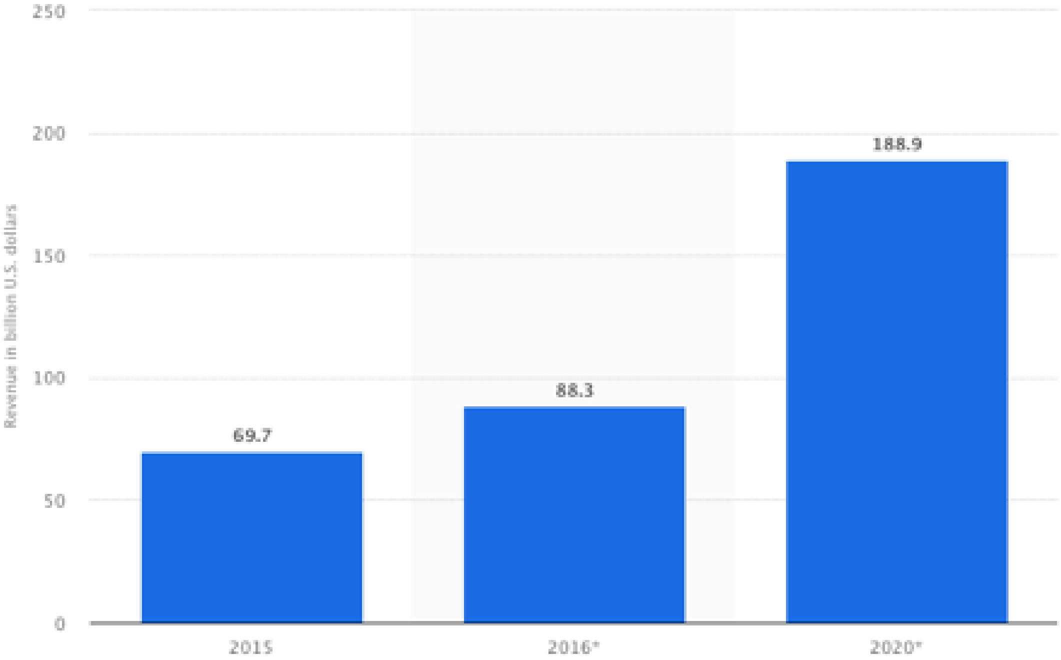
Google Apps Play Store

DAO2702 Positive Vibes

Heng Rui

Introduction to Mobile Application Industry



Growing market, projected to be 188.9billion

Team Positive Vibe

Key Players

Google play

84% market share

Available on the App Store

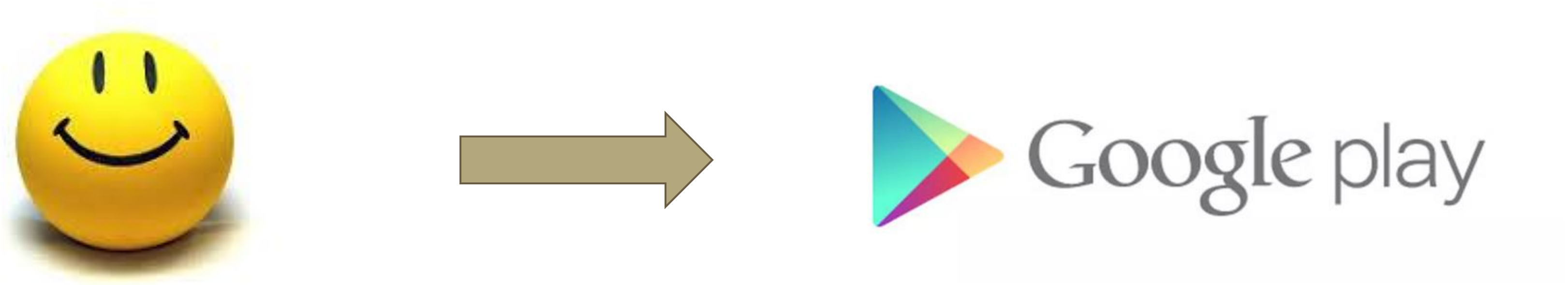
15% market share

Entering Google play market

Reasons:

Cheaper ($25 compared to $99)

Bigger Market



Goals & Vision

Customer Satisfaction

Analysis of Data

10,000 sample dataset

Darren

Definition of Business Problem

How can create a high-rating application on Google Play?

Data Obtained with the following 12 attributes

Category	Type
Reviews	Content Rating
Size	Genres
Installs	Last Updated Version
Price	Current Version
Rating	Android Version

Darren

Modelling Methodologies

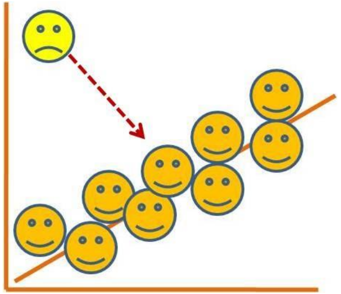
- Multiple Linear Regression Model

Variable	Relationship	Variable Type	Conversion
Y: Rating	Independent Var	Numerical	N.A.
X1: Category	Dependent Variable	Categorical	N.A.
X2: Reviews		Numerical	N.A.
X3: Size		Numerical	N.A.
X4: Installs		Categorical	Numerical
X5: Price		Numerical	Binary (Categorical)

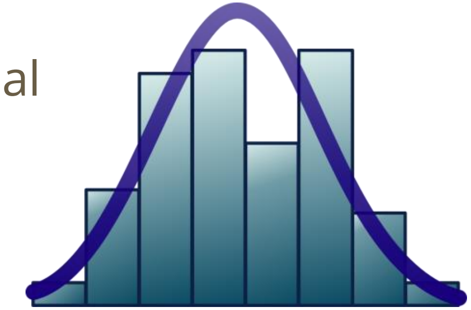
- Hypothesis Testing for Statistical Significance
- Data Visualisation and Estimating Parameters
- Error Minimization and Confidence Intervals

Modelling Assumptions

- Independency of Variables



- Based on Independency, ability to remove outliers in each variable’s set of data



- Based on removing outliers, ability to model Normal Distribution by Central Limit Theorem

Darren

Overview of Data Analysis

- Data Manipulation and Simplification → Darren
- Visualisation of Data: Outliers and Relationship → Su Min
- Simple Linear Regression → Stephenie
- Transformation of Variables → Jun Hyoung
- Multiple Linear Regression and Interactions → Jun Hyoung
- Parameter Estimation through Error Minimization → Kai Xuan
- Evaluation of Confidence Intervals (Limitation) → Kai Xuan
- Interpretation of Results and Solutions for Business Problem → Heng Rui

Darren

Data Manipulation: Removing Unwanted Data

- Removing Unnecessary Columns

Variables	Reasons:
Type	Binary Variable for Price
Content Rating	General App for All
Genres	Elaboration of Category
Last Updated Ver	Only available if Application is launched in the first place.
Current Ver	
Android Ver	

- Removing any rows with NaN
- Conversion of all numerical data types to floats and categorical data types to objects

Darren

Data Manipulation: General Descriptive Statistics

Variable	Mean	S.D.	Category Counts	Distribution
Rating	4.17	0.544	N.A.	Requires Data Visualisation
Category	N.A.	N.A.	33	
Reviews	2.95 x 10^5	1.86 x 10^6	N.A.	
Size	2.29 x 10^7	2.34 x 10^7	N.A.	
Installs	8.41 x 10^6	5.01 x 10^7	19	
Price	1.12	17.4	N.A.	

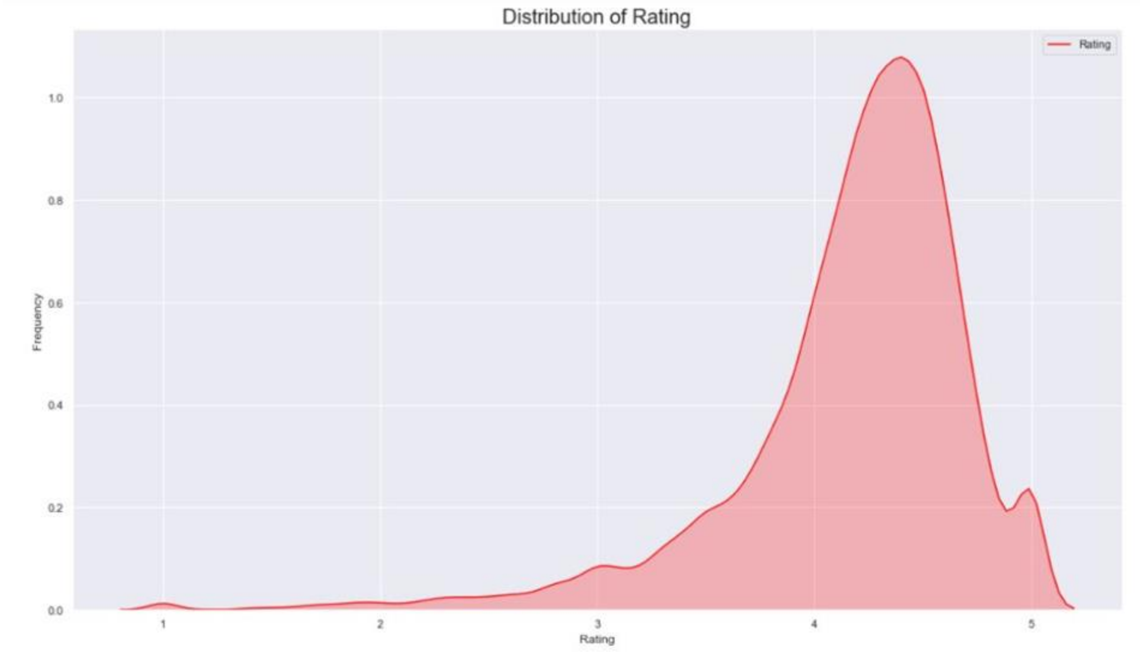
Data Visualisation

- We first conduct data visualization for each X-variable and Y-variable individually
- Determine the range of each variable’s data with low-frequency counts
- Find out the outliers
- See distribution and confidence interval by sample analogue

Suin

Data Visualisation

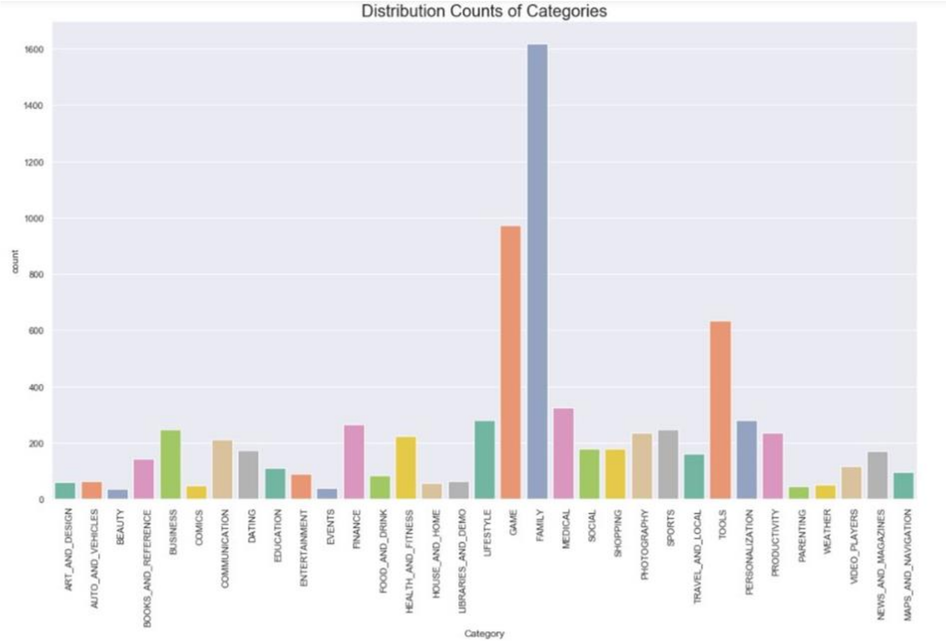
- Distribution of Individual Variables (Y-variable)



Sumin

Data Visualisation

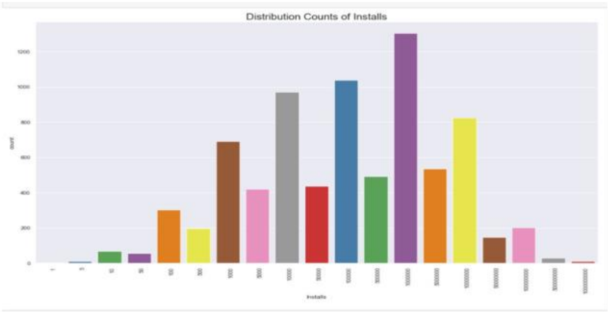
- Distribution of Individual Variables (Categorical X-variables)

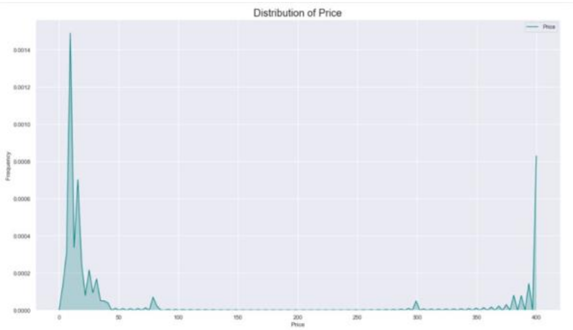


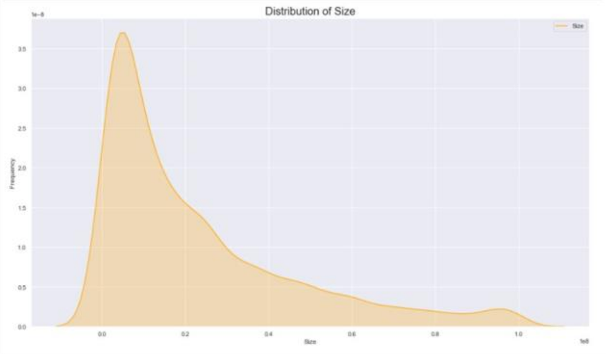
Sumin

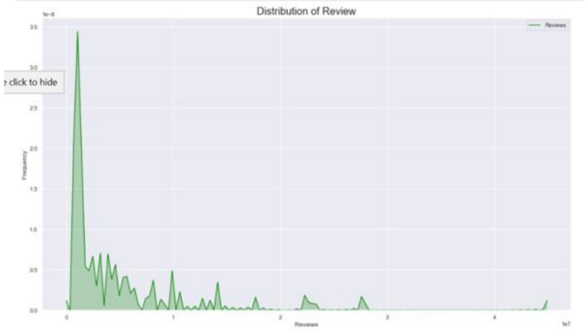
Data Visualisation

- Distribution of Individual Variables (Numerical X-variables)









Data Visualisation

Descriptive Measures:

- Estimate the percentiles at which we obtain about 95% of our data
- Remaining 5% of our data → outliers due to the low-frequency counts

Sumin

Data Visualisation

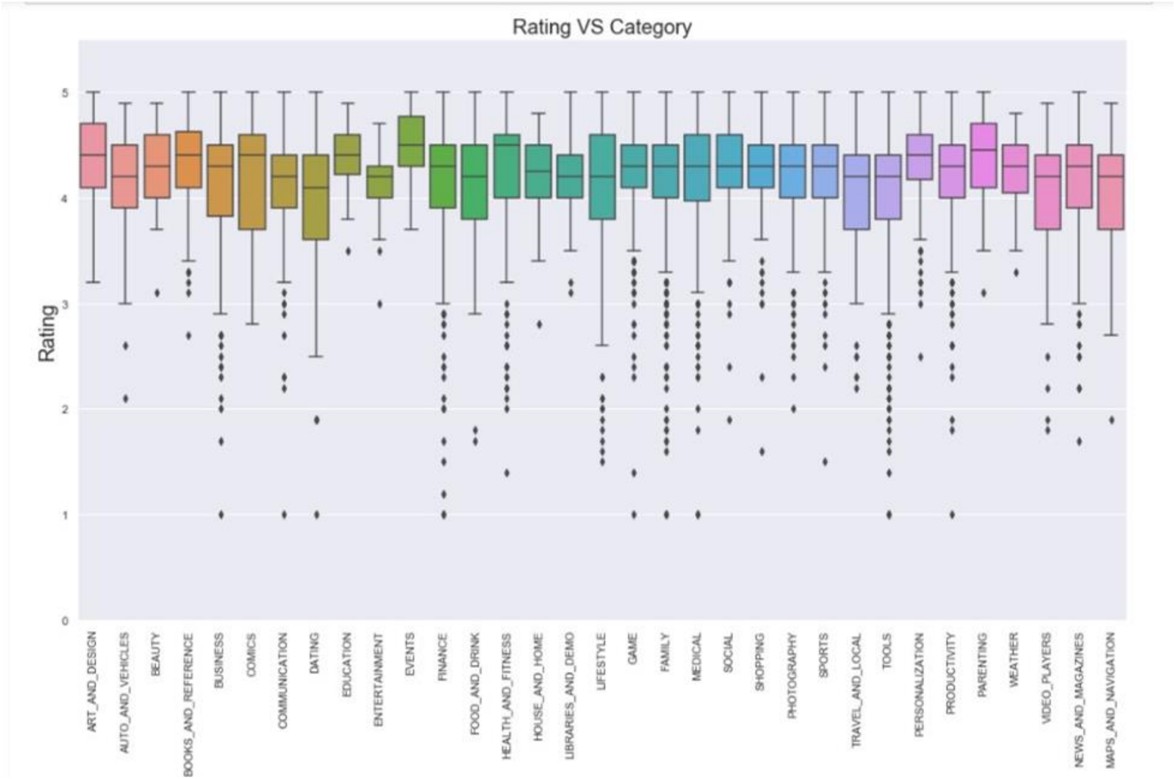
Relationship Analysis

- Scatter plot with histograms → numerical variables
- Side by side boxplots → categorical variables
- See relationships between 2 variables

Sumin

Data Visualisation

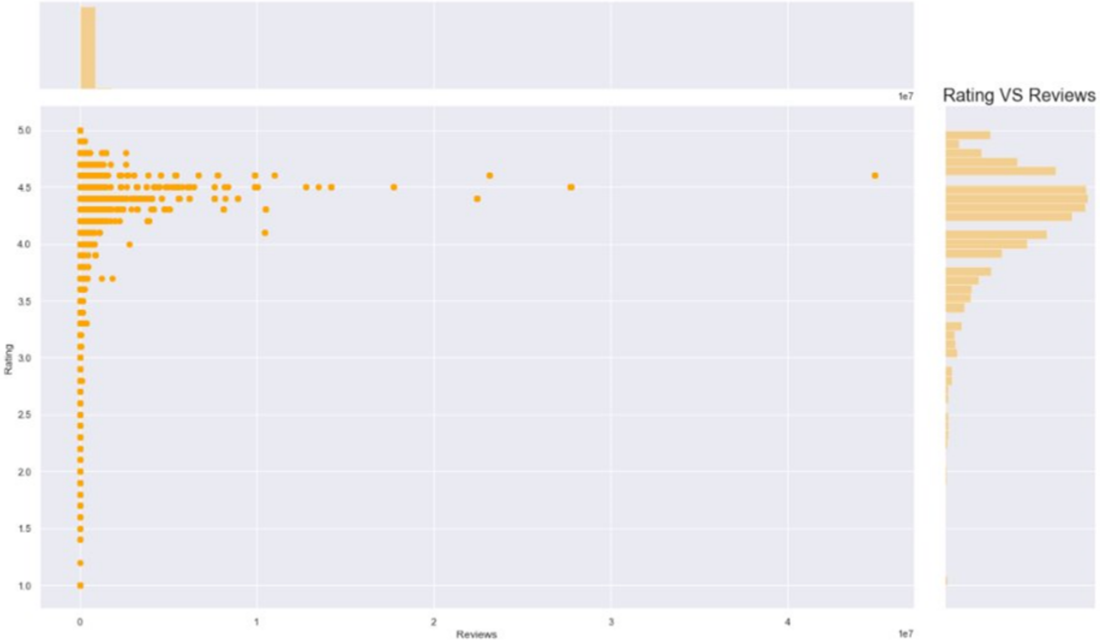
Relationship Analysis



Sumin

Data Visualisation

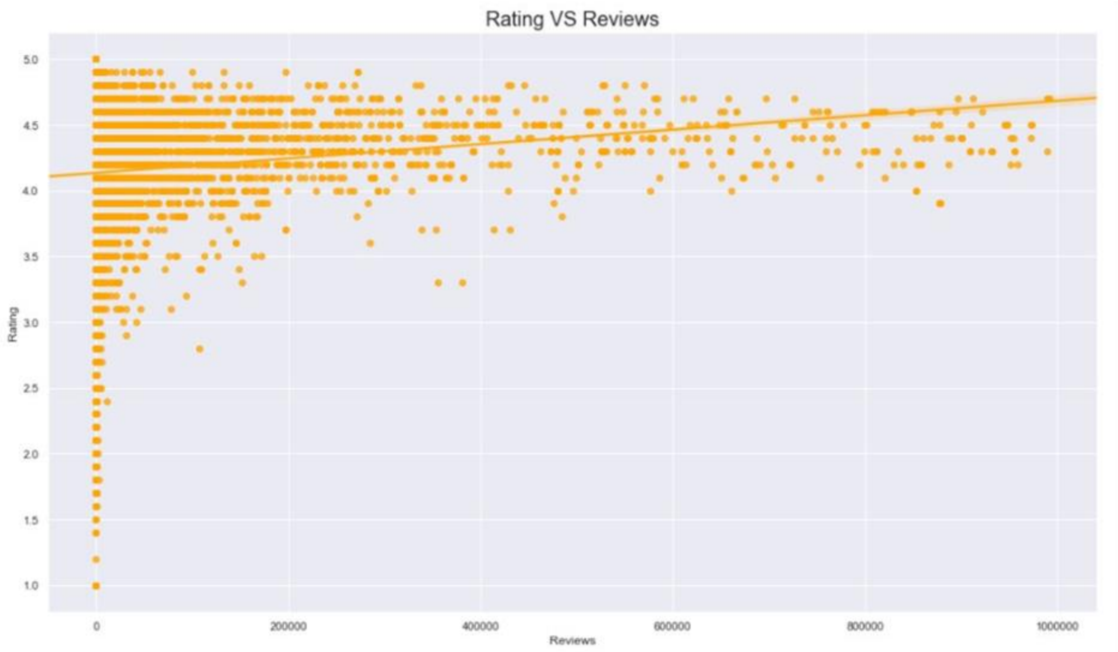
Relationship Analysis



Sumin

Data Visualisation

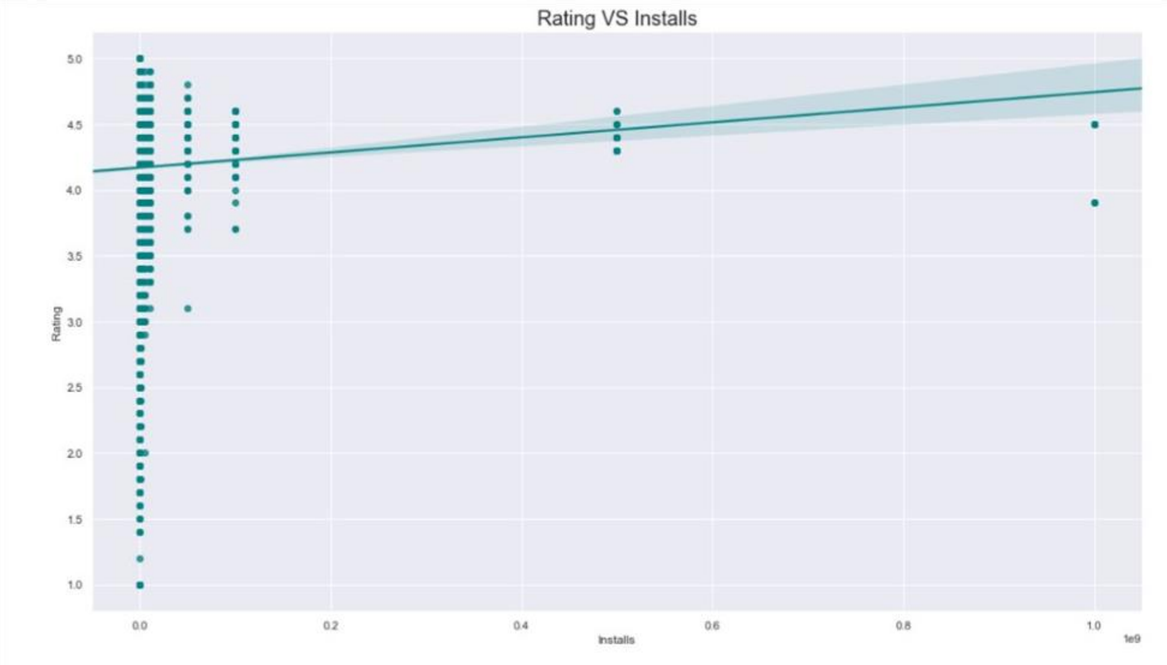
Relationship Analysis



Sumin

Data Visualisation

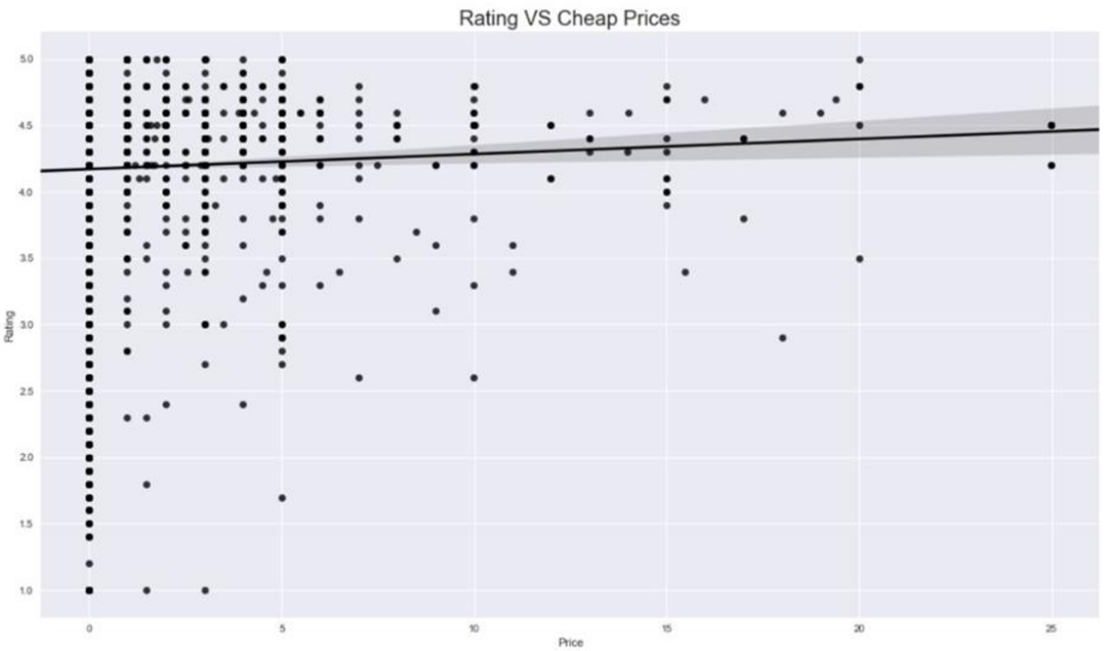
Relationship Analysis



Sumin

Data Visualisation

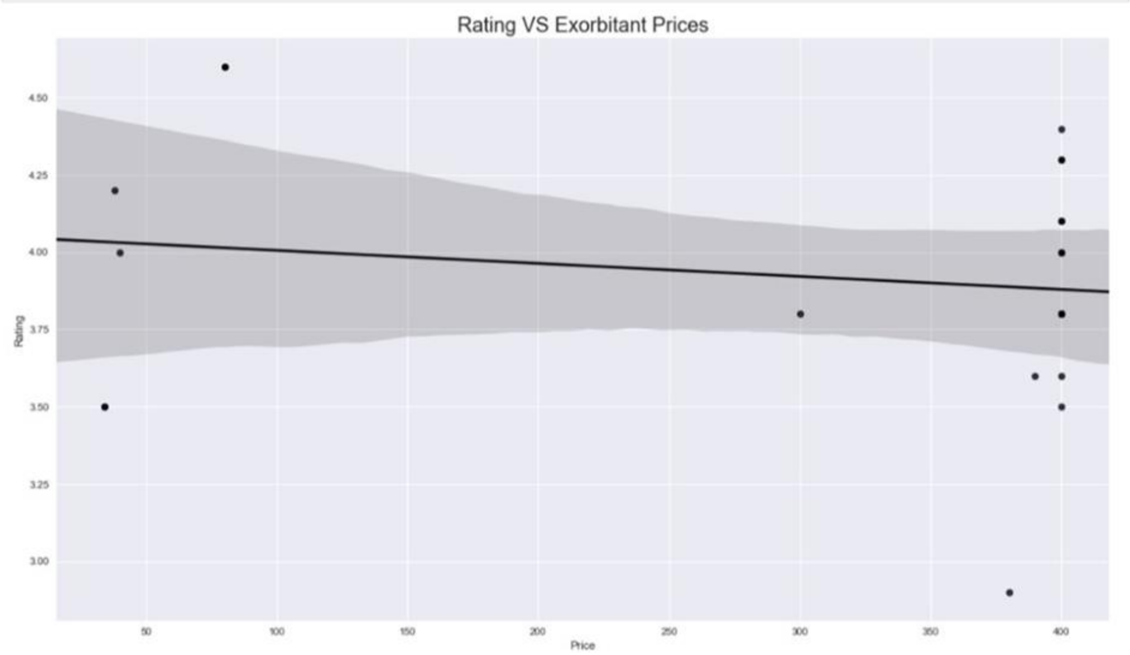
Relationship Analysis



Sumin

Data Visualisation

Relationship Analysis



Sumin

Identifying Outliers

- To **reduce the spread of data** and model distribution **symmetrically about central tendency**
- “Reviews” and “Price” had the greatest range of outliers
- We could not drop the upper range of “Reviews” and upper range of “Price” simultaneously
- Applications that had high numbers of reviews were mostly of low prices, vice versa

Sumin

Choice of Removal of Outliers

- Drop the upper range of “Reviews”
- Exclude Reviews that are more than 1 million to remove the majority outliers → Introduce new applications which could not possibly reach a high number of reviews over a short period of time
- Split the “Price” data into 2 categories → Cheaper applications below $4.99 Vs Expensive applications above $4.99

Stephenie

Simple Linear Regression

X1	Category
X2	Reviews
X3	Size
X4	Installs
X5	Price

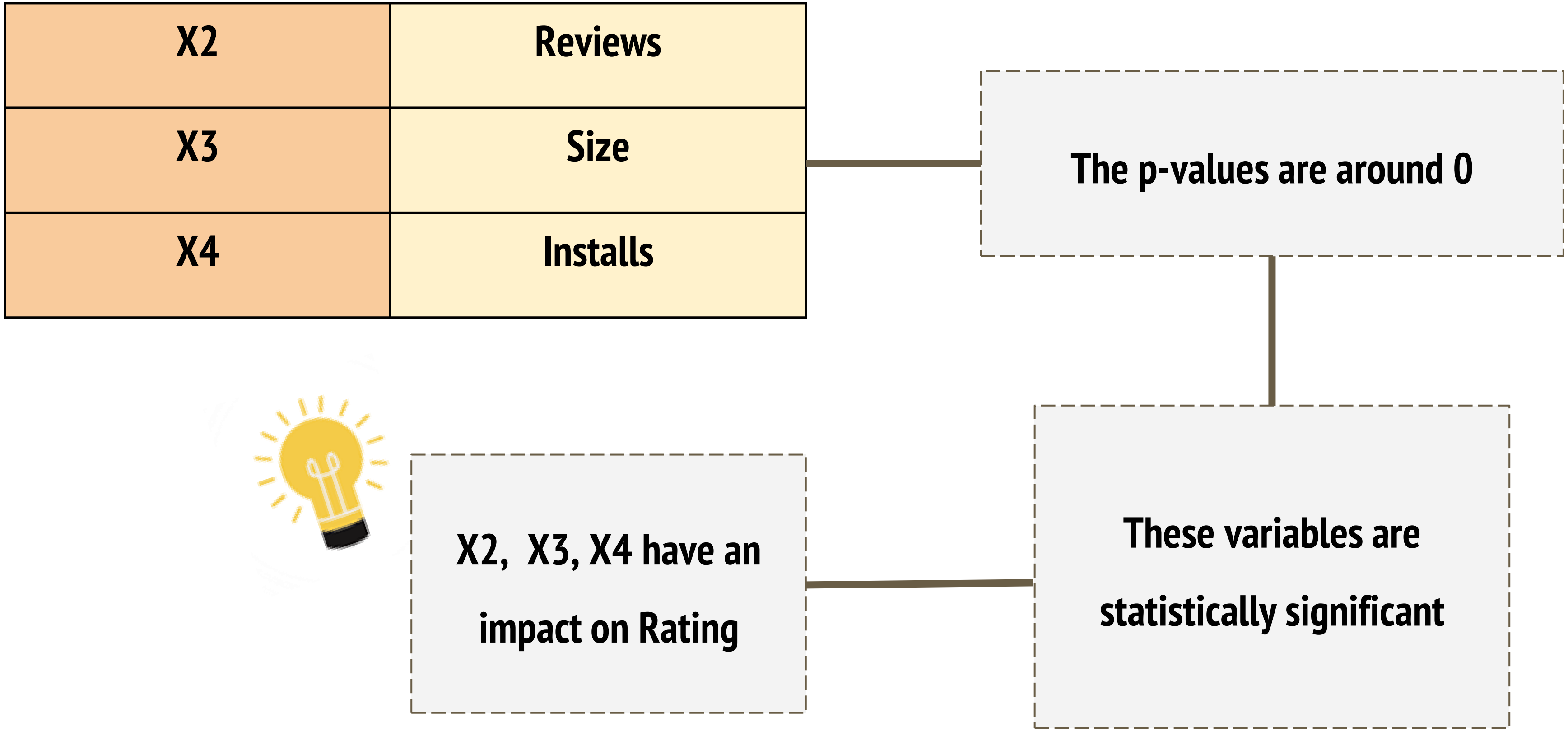
Statistically significant

Determine the linear relationship

X1	Category

- Business
- Entertainment
- Food and Drink
- Map and Navigation
- Medical
- Travel
- News and Magazine
- Dating
- Lifestyle
- Sports
- Comics
- Photography
- Communication
- Family
- Video Players
- Auto and vehicles

20 out of 33 categories are statistically significant and have an impact on Rating

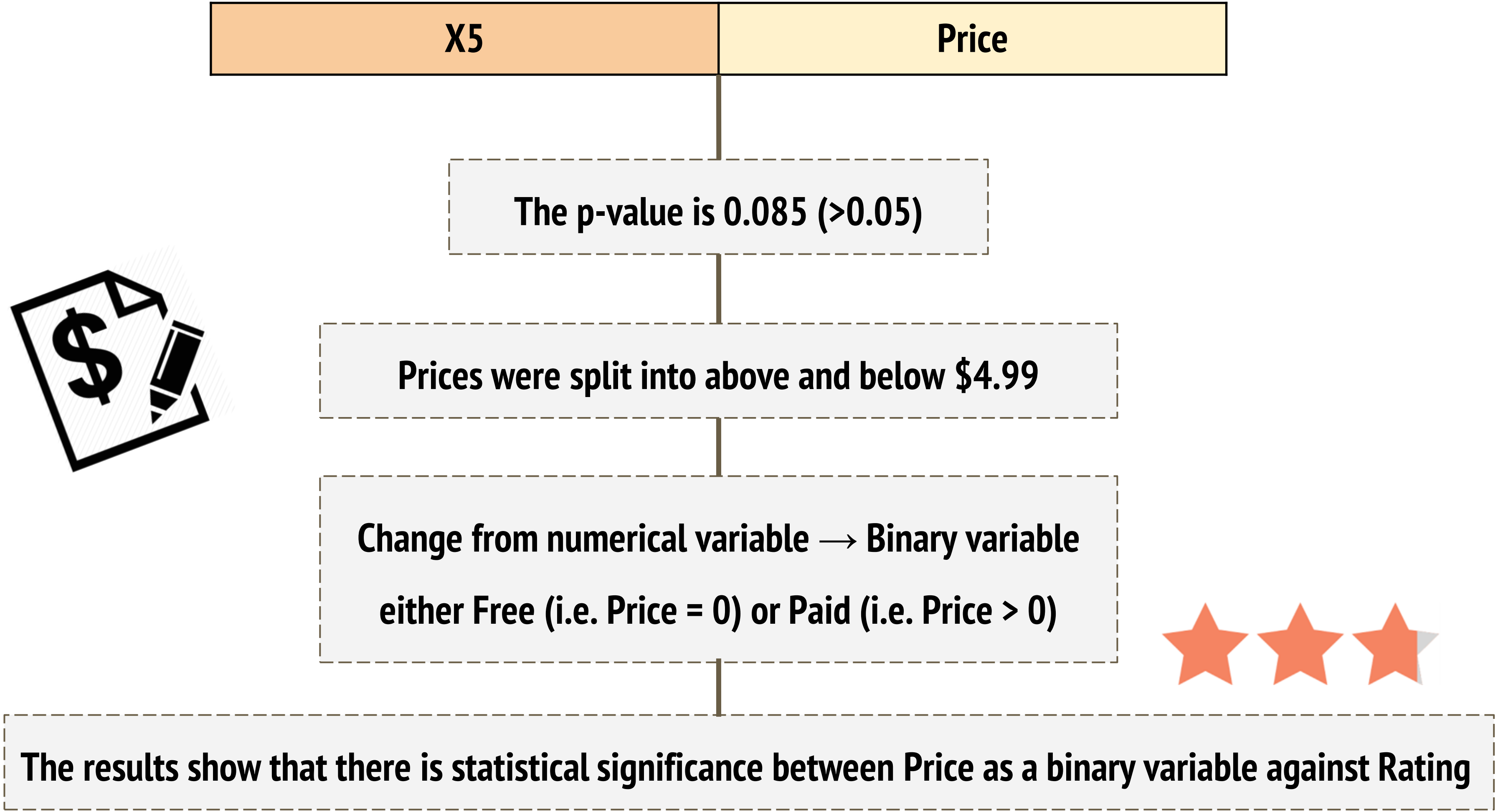


X2	Reviews
X3	Size
X4	Installs

The p-values are around 0

These variables are statistically significant

X2, X3, X4 have an impact on Rating



X5

Price

The p-value is 0.085 (>0.05)

Prices were split into above and below $4.99

Change from numerical variable → Binary variable

either Free (i.e. Price = 0) or Paid (i.e. Price > 0)

The results show that there is statistical significance between Price as a binary variable against Rating

Jun Hyoung

Multiple Linear Regression

- Using multiple linear regression to plot the relationship between Y (Rating) and all the X variables
- Several transformations and interactions to note

Multiple Linear Regression

Transformation:

- Reviews better suit a model of square root function with Ratings
- Logarithmic transformation of Size better suits the function between Size and Rating
- Installs better suits a model of a quadratic function with Ratings

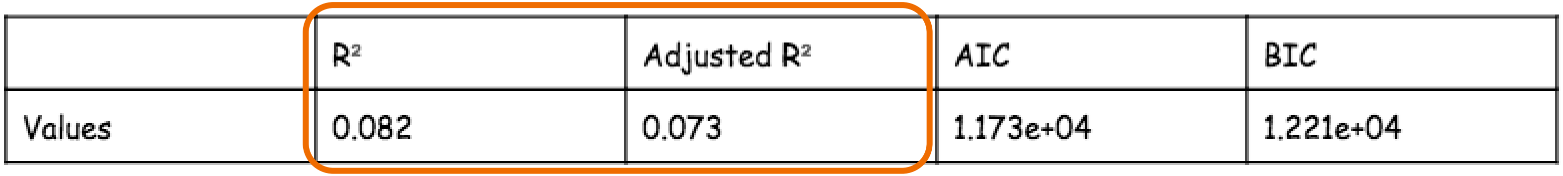
Interaction (By Research):

- Relationship between rating and reviews
- Relationship between price and installs

(Might still have other interactions to consider)

Jun Hyoung

Multiple Linear Regression

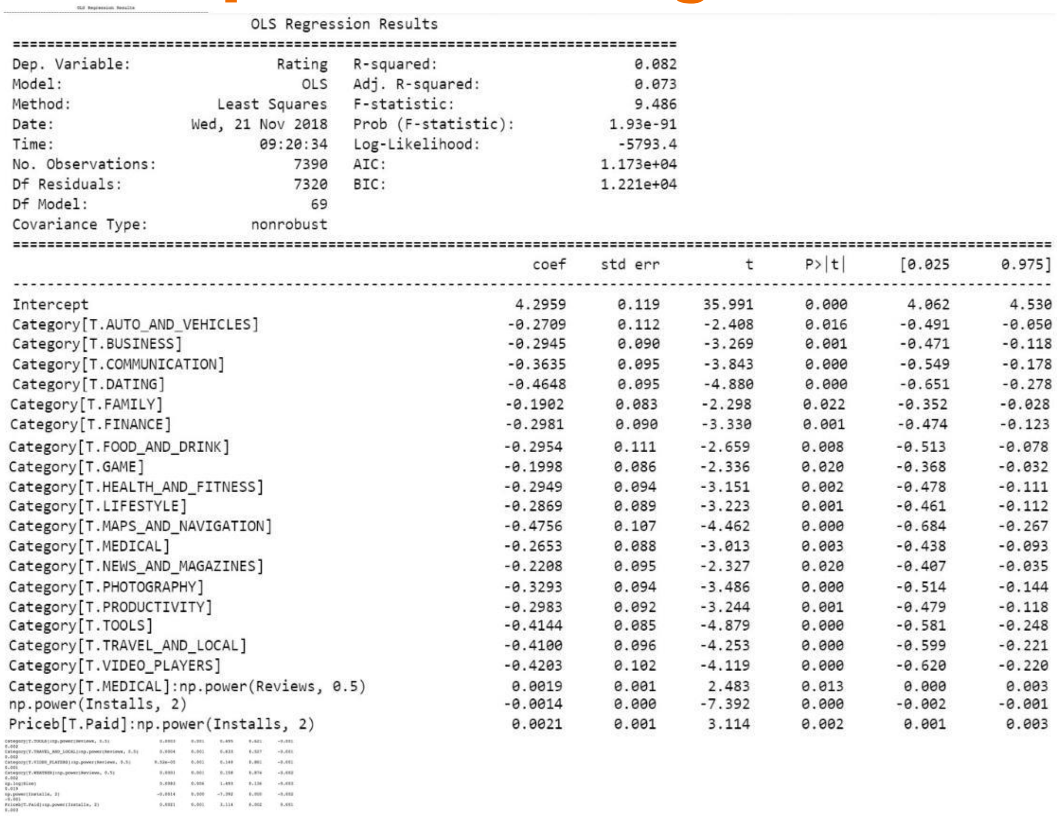


	R^2	Adjusted R^2	AIC	BIC
Values	0.082	0.073	1.173e+04	1.221e+04

- Values suggest that there are very weak relationship between the x variables and rating
- Suggests that none of these x variables affect rating, but possibly some other factor

Jun Hyoung

Multiple Linear Regression



```
OLS Regression Results

                            OLS Regression Results
==============================================================================
Dep. Variable:                 Rating   R-squared:                       0.082
Model:                            OLS   Adj. R-squared:                  0.073
Method:                 Least Squares   F-statistic:                     9.486
Date:                Wed, 21 Nov 2018   Prob (F-statistic):           1.93e-91
Time:                        09:20:34   Log-Likelihood:                -5793.4
No. Observations:                7390   AIC:                         1.173e+04
Df Residuals:                    7320   BIC:                         1.221e+04
Df Model:                          69
Covariance Type:            nonrobust
===============================================================================================================
                                                            coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------------------------
Intercept                                                 4.2959      0.119     35.991      0.000       4.062       4.530
Category[T.AUTO_AND_VEHICLES]                            -0.2709      0.112     -2.408      0.016      -0.491      -0.050
Category[T.BUSINESS]                                     -0.2945      0.090     -3.269      0.001      -0.471      -0.118
Category[T.COMMUNICATION]                                -0.3635      0.095     -3.843      0.000      -0.549      -0.178
Category[T.DATING]                                       -0.4648      0.095     -4.880      0.000      -0.651      -0.278
Category[T.FAMILY]                                       -0.1902      0.083     -2.298      0.022      -0.352      -0.028
Category[T.FINANCE]                                      -0.2981      0.090     -3.330      0.001      -0.474      -0.123
Category[T.FOOD_AND_DRINK]                               -0.2954      0.111     -2.659      0.008      -0.513      -0.078
Category[T.GAME]                                         -0.1998      0.086     -2.336      0.020      -0.368      -0.032
Category[T.HEALTH_AND_FITNESS]                           -0.2949      0.094     -3.151      0.002      -0.478      -0.111
Category[T.LIFESTYLE]                                    -0.2869      0.089     -3.223      0.001      -0.461      -0.112
Category[T.MAPS_AND_NAVIGATION]                          -0.4756      0.107     -4.462      0.000      -0.684      -0.267
Category[T.MEDICAL]                                      -0.2653      0.088     -3.013      0.003      -0.438      -0.093
Category[T.NEWS_AND_MAGAZINES]                           -0.2208      0.095     -2.327      0.020      -0.407      -0.035
Category[T.PHOTOGRAPHY]                                  -0.3293      0.094     -3.486      0.000      -0.514      -0.144
Category[T.PRODUCTIVITY]                                 -0.2983      0.092     -3.244      0.001      -0.479      -0.118
Category[T.TOOLS]                                        -0.4144      0.085     -4.879      0.000      -0.581      -0.248
Category[T.TRAVEL_AND_LOCAL]                             -0.4100      0.096     -4.253      0.000      -0.599      -0.221
Category[T.VIDEO_PLAYERS]                                -0.4203      0.102     -4.119      0.000      -0.620      -0.220
Category[T.MEDICAL]:np.power(Reviews, 0.5)                0.0019      0.001      2.483      0.013       0.000       0.003
np.power(Installs, 2)                                    -0.0014      0.000     -7.392      0.000      -0.002      -0.001
Priceb[T.Paid]:np.power(Installs, 2)                      0.0021      0.001      3.114      0.002       0.001       0.003
Category[T.TOOLS]:np.power(Reviews, 0.5)                  0.0003      0.001      0.495      0.621      -0.001
0.002
Category[T.TRAVEL_AND_LOCAL]:np.power(Reviews, 0.5)       0.0004      0.001      0.633      0.527      -0.001
0.002
Category[T.VIDEO_PLAYERS]:np.power(Reviews, 0.5)        9.52e-05      0.001      0.140      0.881      -0.001
0.001
Category[T.WEATHER]:np.power(Reviews, 0.5)                0.0001      0.001      0.158      0.874      -0.002
0.002
np.log(Size)                                              0.0083      0.006      1.493      0.136      -0.003
0.019
np.power(Installs, 2)                                    -0.0014      0.000     -7.392      0.000      -0.002
-0.001
Priceb[T.Paid]:np.power(Installs, 2)                      0.0021      0.001      3.114      0.002       0.001
0.003
```

- Lot of extreme P values
- Filtered those with P values less than 0.5

Evaluation of Relationships

- Estimation of parameters by coercing a 95% significant data range through sample analogue method and through observation from data visualisation may be ineffective
- Evaluate our **Confidence Intervals** (CI) of our x-variable data through **Error Minimisation** method

Kai Xuan

Evaluation of Relationships

Error-Minimisation Model: $\boldsymbol{Y_i = \mu + \epsilon_i}$

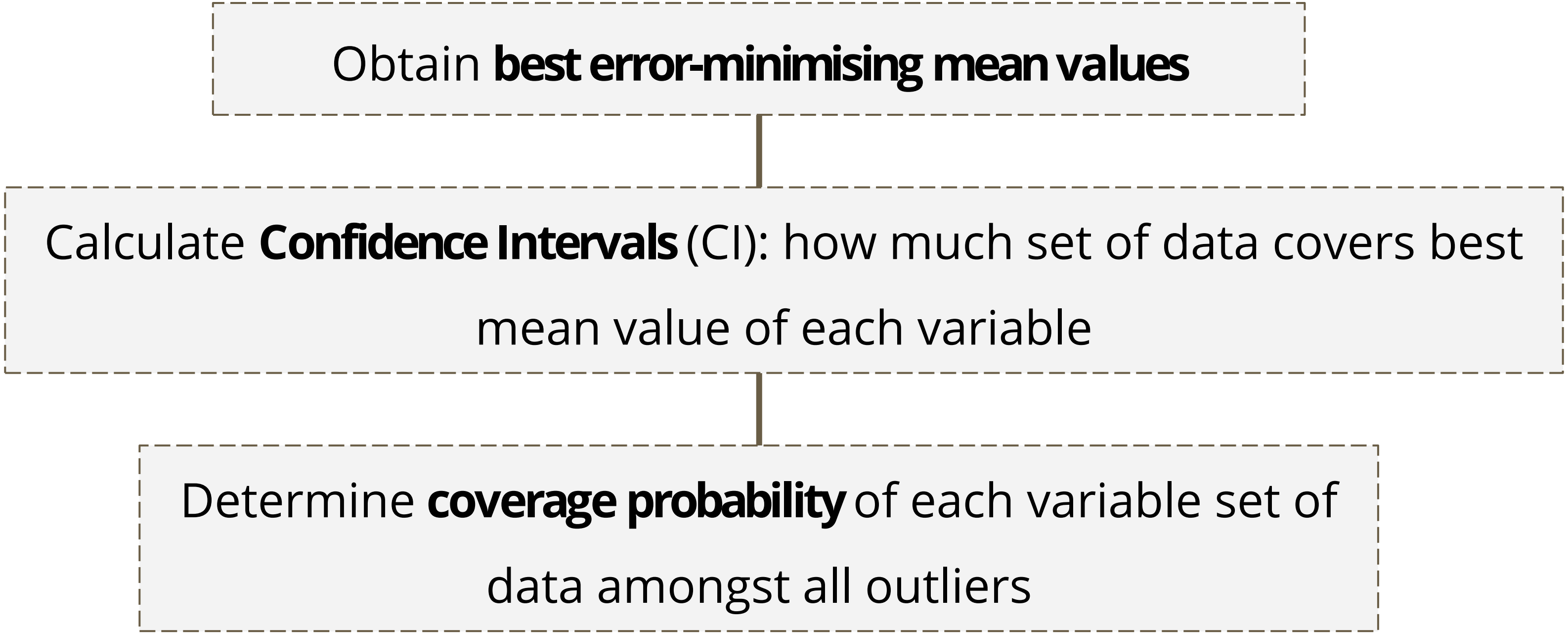
where i = application index in the sample, and i = 1, 2, ... 7390

μ = population parameter to represent averages of various variables (i.e. reviews, size, installs, price)

ϵ = application i's error, the deviation from the average

Evaluation of Relationships

Error-Minimisation Method



Obtain **best error-minimising mean values**

Calculate **Confidence Intervals** (CI): how much set of data covers best mean value of each variable

Determine **coverage probability** of each variable set of data amongst all outliers

Evaluation of Relationships

Coverage Probability

Variable	Best New Mean	Coverage Probability of Data (given best mean value)
Reviews	51,759.9484	0.9467
Size	21,633,404.1819	0.9497
Installs	9.7393	0.9561
Price	1.1910	0.9515
Rating	4.1620	Not Needed

Evaluation of Relationships

Statistical Significance

- At 95% confidence level, all variables are **statistically significant**
- Reject null hypothesis and conclude that all variables have a relationship with Rating

Kai Xuan

Limitations to Modelling Assumptions

- Since we have removed Reviews that are above 1 million, associated values of other x-variables might be in the original calculated confidence intervals of the other x-variables
- Old confidence intervals will have many missing data points
- Calculate coverage probabilities to ensure that our sample data is significant at 95% confidence level
- Hypothesis testing will be based on new set data and its associated confidence intervals

Hypothesis Testing

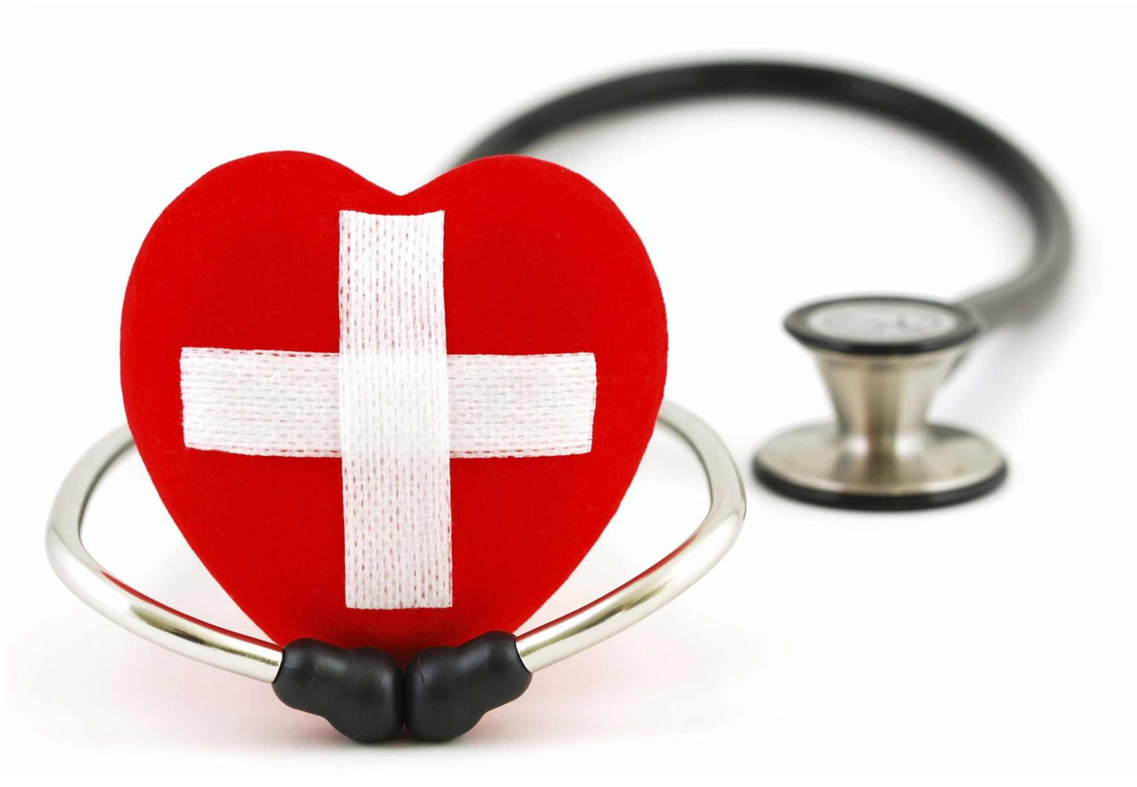
Variable 1	Variable 2	Correlation Coefficient
Reviews	Size	0.2349556
Reviews	Installs	0.50302441
Reviews	Price	-0.02429811
Size	Installs	0.27068609
Size	Price	-0.0242148
Price	Installs	-0.05762671

Summary Table of Final Regression Model

The combination of factors to give highest Rating:

Variable Mix	p-value	Top Ranked Coefficients
Price [Paid]	0.002	0.0021
Category: Medical with Reviews	0.013	0.0019

Positive Vibe High Rating App:



Thank You