INDIVIDUAL COURSEWORK COVERSHEET

MSc. Business Analytics

Module : Statistics and Econometrics, MANM526

Date : 2nd Jan / 2024

Name : Mohamed Shahil Shahul Hameed

URN No : 6809743

Words : 3150

STATISTICS AND ECONOMETRICS

INTRODUCTION:

In this report, we explore the effectiveness of competitive strategies for app success in the Google Play store, focusing on a dataset of paid apps from June 2023. The study addresses the challenge of standing out in a highly competitive app market where users must pay upfront to download and install apps. By analysing many factors such as app revenue, ratings, price, and monetization strategies, we aim to find key elements that contribute to achieving higher downloads and revenue, offering insights into successful strategies in this dynamic digital marketplace.

DATA:

The dataset for this study includes variables from a sample of paid apps on the Google Play store, captured in June 2023. It features key indicators such as app revenue in USD, user ratings, app price, and monetization strategies, including in-app purchases and advertisements. Added variables include the target age group, the number of languages the app is available in, the app's file size, and its main category. These elements are complemented by general information like app name, developer, release date, and version, supplying a comprehensive view of the factors influencing app success in a competitive digital market.

VARIABLE:

Dependent variable: The variable like ‘revenue' reflecting the financial success of the apps is dependent variable.

Independent variables: The variable like 'price', 'size', 'rating', 'age_target', and 'mon_strategy_num', which are factors hypothesized to influence app revenue.

Control variables: The variable like 'num_langs','category_num', which account for other influences on revenue but are not the primary focus of the analysis. These variables collectively allow for a nuanced understanding of what drives financial success in the app market.

MODEL:

Ordinary Least Square (OLS) regression is run for the following baseline model in this study. The standard error is robust to expected heteroscedasticity in all models.

log_revenueit = β0 + β1 × log_priceit + β2 × monetization_strategiesit + β3 × ratingit + βj × age_targetit + βk × main_categoryit + βl × num_langsit + ε it

DESCRIPTIVE ANALYSIS:

Summary statistics and a correlation matrix are used to conduct a descriptive analysis and assess relationships among various variables. Summary statistics for the full sample are detailed in Tables 1-7, excluding categorical variables.

Table 1: Summary for the entire dataset

Variable	Obs	Mean	std. dev.	Min	Max
product_id	0				
name	0				
developer	0				
version	0				
release_date	0				
devices	0				
active	0				
price	3170	5.313785	7.47637	0.05	199.99
is_paid	0				
size	2785	5.35E+07	1.15E+08	1.02E+03	1.61E+09
rating	3173	4.1237	0.606698	1	5
num_langs	3145	12.56757	19.33077	1	69
main_categ~y	0				

monetizati~s	0				
age_target	0				
revenue	3173	7967386	92100000	1000	4.89E+09
log_revenue	3173	13.24946	2.27543	6.907755	22.31113
log_price	3170	1.310516	0.825423	-1.89712	5.298767
mon_strate~m	3142	1.346595	0.734288	1	3
category_num	3145	1.47E+01	6.963364	1.00E+00	3.10E+01
category_num	3145	14.73831	6.963364	1	31

The table 1 shows the analysis of the Google Play store apps reveals a broad price range from $0.05 to $199.99, reflecting diverse monetization approaches. App sizes vary widely, showing a mix of simple and complex offerings. With an average rating of 4.12, most apps are well-received by users. The availability of apps in multiple languages shows an effort to cater to a global audience. Revenue disparities are significant, highlighting varying levels of success among the apps. These insights are crucial for understanding the factors influencing app performance in this competitive market.

Correlation Matrix:

Table 2: correlation matrix

	(1)				
	log_revenue	rating	log_price	num_langs	size
log_revenue	1				
rating	0.202^{***}	1			
log_price	0.0342	0.0396^{*}	1		
num_langs	0.237^{***}	0.110^{***}	0.0173	1	
size	0.0871^{***}	0.00675	0.128^{***}	0.0129	1
Observations	2757				

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2 correlation matrix delineates the relationships between app attributes and financial performance on Google Play. A moderate positive correlation between ratings and revenue suggests that user satisfaction can influence financial gains. The stronger correlation with the number of languages shows a wider audience reach correlates with higher revenue. Although price and app size show positive correlations with revenue, these relationships are less pronounced, implying that soaring prices and large sizes don't necessarily equate to higher earnings.

Summary Statistics:

Table 3: Summary statistics for Gaming, Non-gaming apps and Total variables

	(1)				
	n	mean	min	max	sd
Gaming Apps	1359	13.90913	7.600903	22.31113	2.2655
	1359	1.279777	-0.9162907	3.499231	0.734427
	1359	4.192053	1.6	5	0.5062163
	1359	10.5894	1	69	15.07387
	1185	8.38E+07	82944	1.61E+09	1.57E+08
Non-Gaming Apps	1786	12.7375	6.907755	19.70641	2.144445
	1783	1.33528	-1.89712	5.298767	0.8892422
	1786	4.074076	1	5	0.6664887
	1786	14.07279	1	69	21.91037
	1573	3.08E+07	1024	1.18E+09	5.82E+07
Total	3145	13.24378	6.907755	22.31113	2.272605
	3142	1.311274	-1.89712	5.298767	0.8261828
	3145	4.125056	1	5	0.6052268
	3145	12.56757	1	69	19.33077
	2758	5.35E+07	1024	1.61E+09	1.15E+08

Table 3 highlights a segment of the Play store with distinct characteristics of Gaming and Non-Gaming Apps : Gaming Apps have a higher average log revenue of 13.9091 shows notable profitability, while a slightly lower average log price of 1.2797 suggests a balance between affordability and revenue. These apps enjoy higher user ratings of 4.1920 and are available in fewer languages (10.5894), focusing on specific audiences. Their larger size (83800 KB) shows more extensive features or content. This analysis shows varied strategies in the app market, from global reach to niche focus.

Non-gaming apps in the Play store generally have lower revenue (average log revenue of 12.7375), but slightly higher prices (average log price of 1.2797) compared to gaming apps. They receive favourable user ratings (4.0740) and cater to a broader audience with availability in more languages (average 14.0727). These apps tend to be smaller in size (average 248800 KB), focusing on specific functionalities or niches. This data highlights the diverse strategies and user reach of non-gaming apps in the app market.

ANNOVA TEST:

Table 4: Revenue differences between game and non-game apps using ANOVA.

Number of jobs = 3,145 R-squared = 0.1193
Root MSE = 2.14305 Adj R-squared = 0.1108

Source	Partial SS	df	MS	F	Prob>F
Model	1936.4	30	64.5466	14.05	0
category_~m	1936.4	30	64.5466	14.05	0
Residual	14301.5	3114	4.59265		
Total	16237.9	3144	5.16473		

The ANOVA analysis in Table 4 writes down significant revenue differences between gaming and non-gaming apps. With an F-statistic of 14.05 and $p < 0.0001$, the impact of app category on revenue is statistically significant. The model explains about 11.93% of the variance in revenue, as shown by the R-squared value, with an adjusted R-squared of 0.1108. This highlights the substantial role of app category in financial success, while acknowledging the presence of other influential factors not covered in this analysis.

Two Sample T-test with equal variances:

Table 5: Conducting a T-test for revenue differences between game and non-game apps.

Group	Obs	Mean	Std. errs.	Std. dev.	[95%
0	1814	12.75524	0.050595	2.15491	12.65601
1	1359	13.90913	0.061455	2.2655	13.78858
Combined	3173	13.24946	0.040395	2.27543	13.17025
diff	-1.1539	0.079034			-1.30885

```
diff = mean(0) - mean(1)                                  t = -14.6000
H0: diff = 0                                Degrees of freedom =  3171

Ha: diff < 0              Ha: diff != 0              Ha: diff > 0
Pr(T < t) = 0.0000       Pr(|T| > |t|) = 0.0000       Pr(T > t) = 1.0000
```

The two-sample t-test reveals that gaming apps have a significantly higher mean log revenue (13.909) than non-gaming apps (12.755), with a substantial difference of -1.154. This statistically significant disparity, underscored by a t-statistic of -14.60 and a p-value much less than 0.0001, shows that gaming apps tend to generate more revenue. High confidence intervals for both groups further confirm these results. This highlights different revenue trends between game and non-game apps in the app marketplace.

EXPLORATORY ANALYSIS:

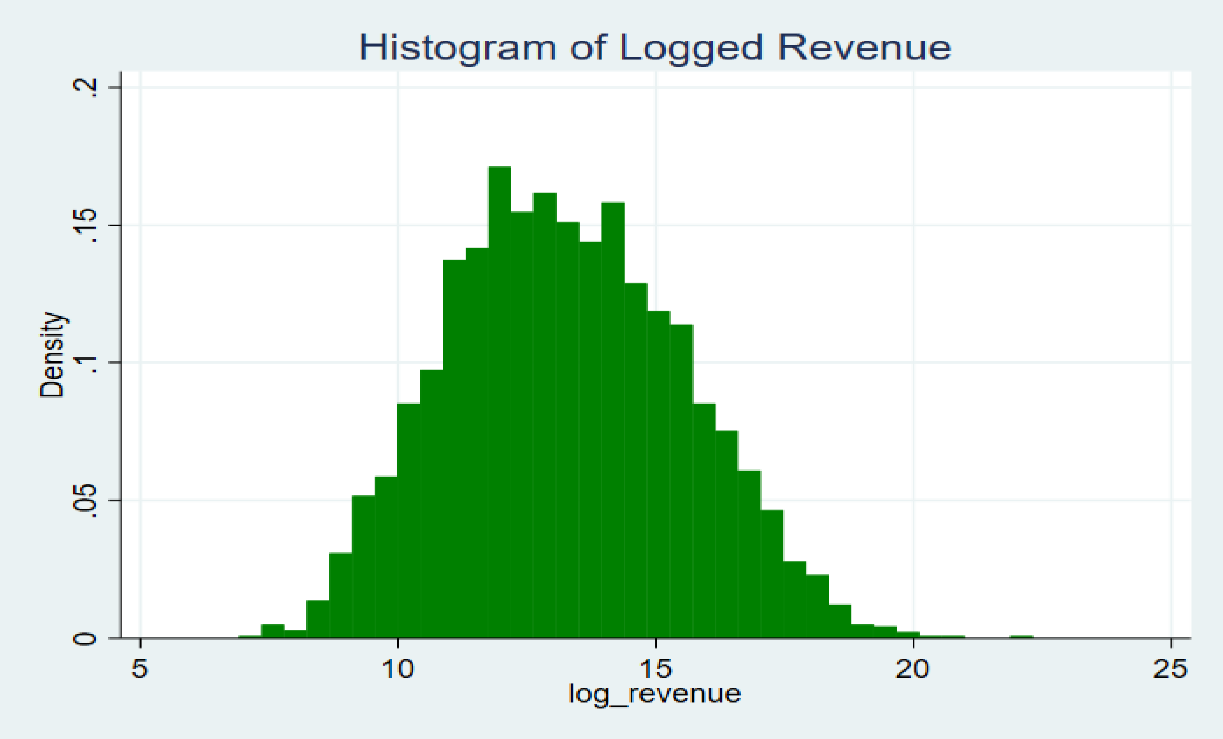

Figure 1: Histogram of Logged Revenue

The histogram titled "Histogram of Logged Revenue" displays the distribution of logarithmically transformed revenue values. The x-axis stands for these log_revenue values, while the y-axis shows their density or relative frequency. The graph reveals a concentration of values around 10, indicating it's the most common log_revenue level. The distribution is right-skewed, suggesting fewer instances of higher revenue values. This suggests a wide range in the original revenue figures, narrowed down post-transformation.

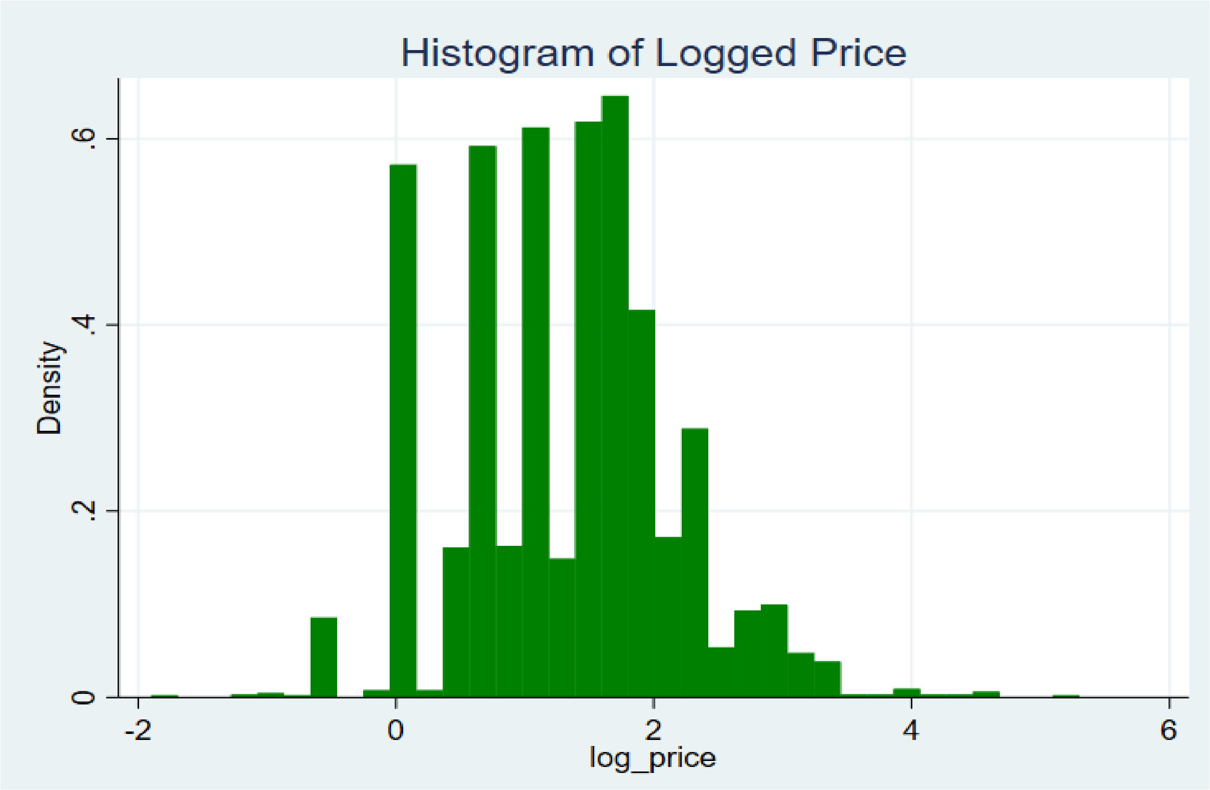

Figure 2: Histogram of Logged Price

The histogram shows a right-skewed distribution of logged app prices, with a quick drop in frequency as prices rise and a dense concentration of apps in the lower price range. The bigger bars on the left suggest that most apps are inexpensive. The small population bars to the right indicate that expensive apps are less common. This pattern indicates that premium-priced apps are uncommon in the app market, which is primarily made up of more reasonably priced apps.

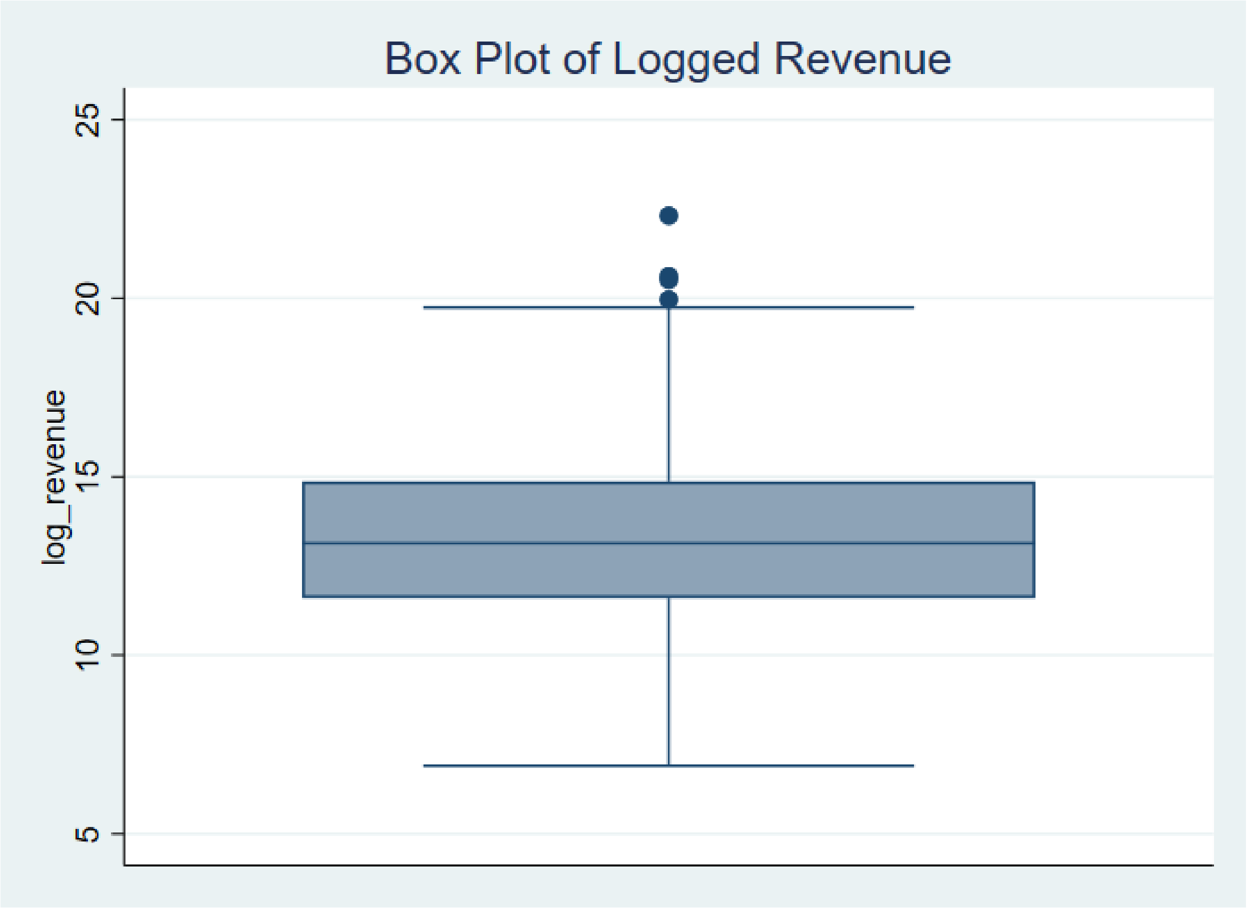

Figure 3: Box plot of Logged Revenue

The "Box Plot of Logged Revenue" graphically stands for the distribution of logarithmically transformed revenue. The median, shown as a horizontal line within the box, writes down the central tendency, while the box depicts the interquartile range (middle 50% of the data). The whiskers extend to show the range of typical values, and points beyond these are potential outliers. This plot highlights the median log_revenue, its spread, variability, and any skewness in the distribution.

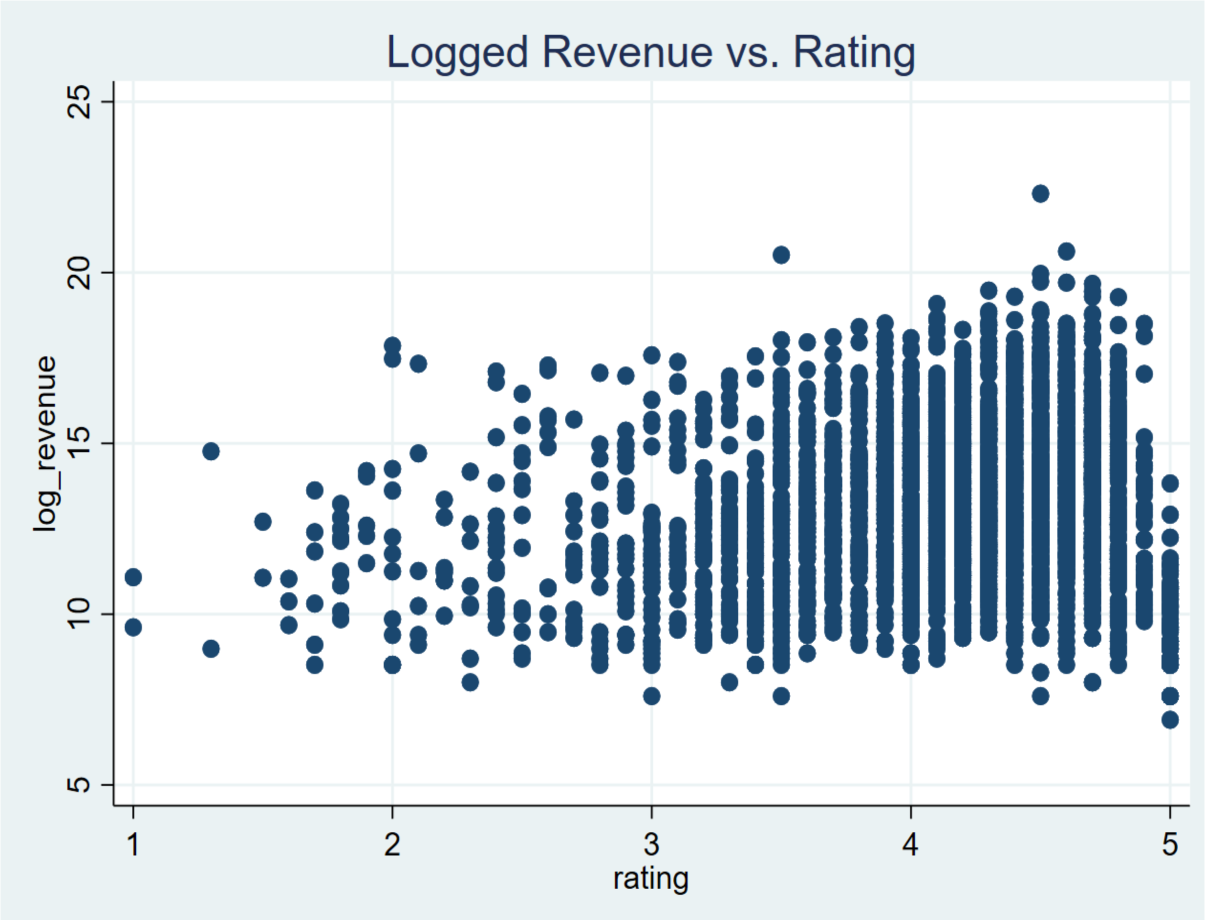

Figure 4: Logged Revenue vs Rating

The scatter plot titled "Logged Revenue vs. Rating" shows the relationship between logged revenue and ratings. On the x-axis is log_revenue, and on the y-axis is rating. The scattered points show a general, but not perfect, positive correlation between higher revenues and higher ratings. The plot also shows variability and some potential outliers in the data.

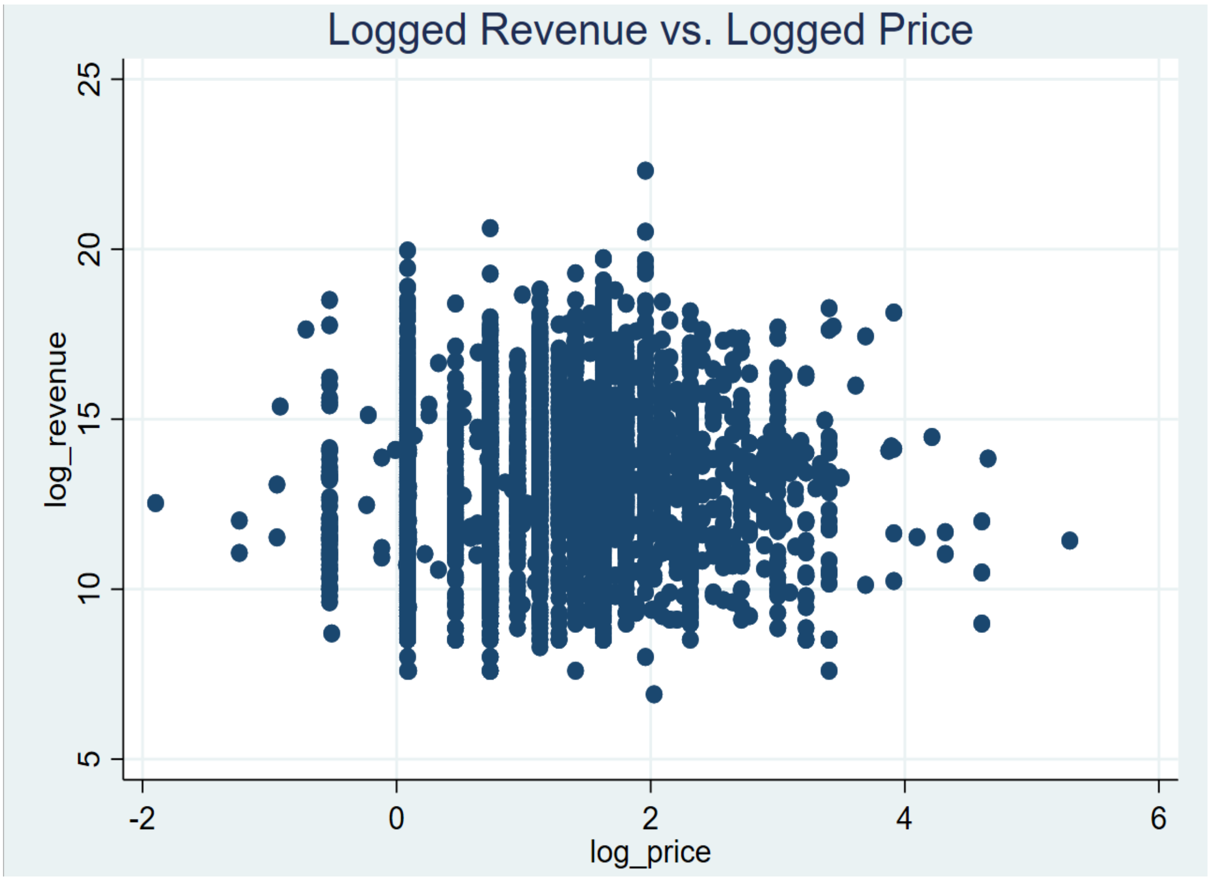

Figure 5: Logged Revenue vs Logged price.

The scatter plot "Logged Revenue vs. Logged Price" displays each observation's log_price and log_revenue. It shows a positive, non-linear correlation between log_revenue and log_price, with variability and potential outliers. The non-linear relationship suggests complexity beyond a straightforward linear association.

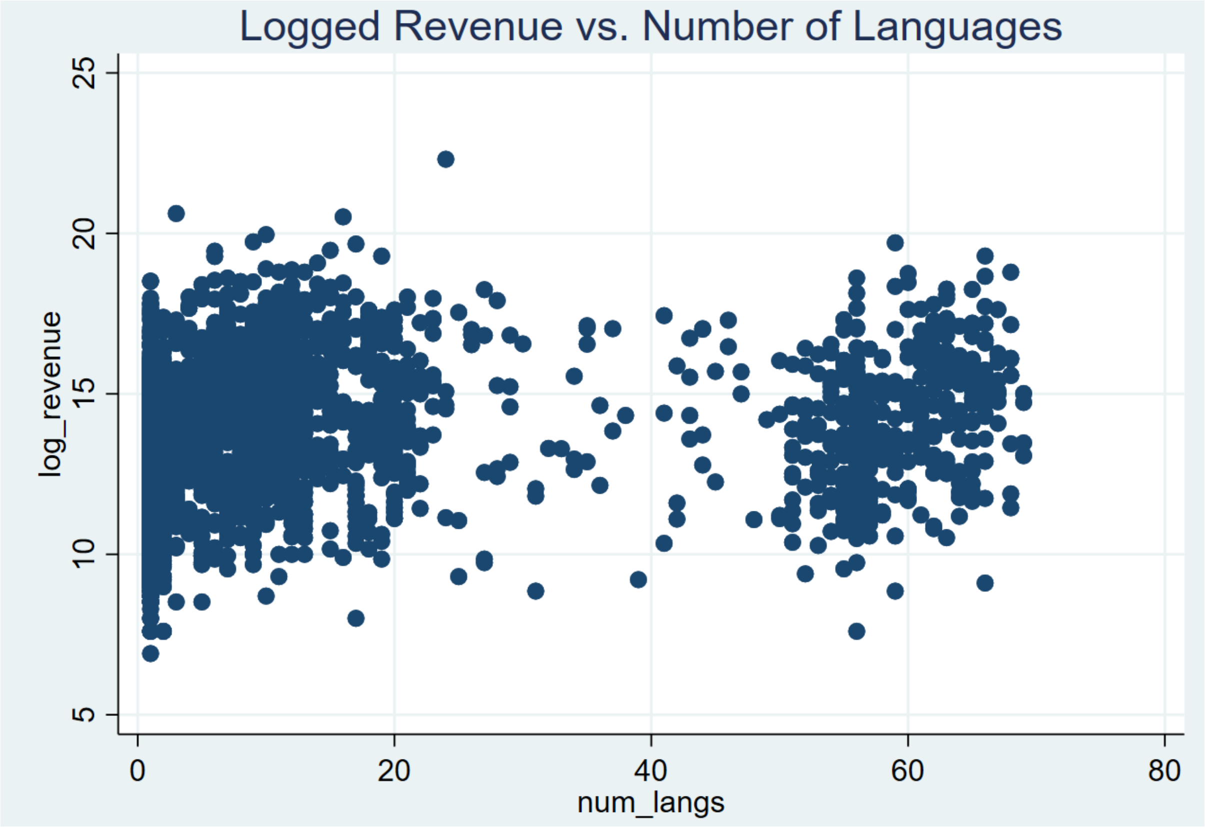

Figure 6: Logged Revenue and the Number of Languages available in an app.

The scatter plot presented visualizes the relationship between revenue, transformed using a logarithm, and the number of languages. Transformations like the natural logarithm are often utilized to standardize data distributions and diminish asymmetry, thereby clarifying real trends and enhancing the robustness of the statistical evaluations. The plot exhibits bifurcation into two clusters, which could imply the presence of a hidden categorical variable influencing the association between the number of languages and the revenue. The graph does not exhibit a uniform linear relationship across the spectrum of 'num_langs', indicating that increments in the number of languages do not uniformly correspond to increased revenue when assessed through the lens of logarithmic revenue.

MAIN REGRESSION

BASELINE MODEL:

Table 6 : (Model 1) Running OLS regression with specified independent variables & (Model 2) Implementing OLS regression with interaction terms.

	(Model 1)	(Model 2)
rating	0.521*** (0.061)	0.489*** (0.065)
log_price	0.149** (0.046)	0.151** (0.046)
Everyone	0.000 (.)	0.000 (.)
Everyone 10+	0.795*** (0.148)	0.805*** (0.148)
Mature 17+	0.787*** (0.231)	0.767*** (0.231)
Teen	0.782*** (0.119)	0.790*** (0.119)
paid	0.000 (.)	0.000 (.)
paid;ads	0.455* (0.202)	3.875* (1.610)
paid;inapp	1.173*** (0.100)	-0.747 (0.785)
num_langs	0.024*** (0.002)	0.024*** (0.002)
Art & Design	Included	Included
rating		0.000 (.)

paid # rating		0.000
		(.)
paid;ads # rating		-0.838*
		(0.391)
paid;inapp # rating		0.461*
		(0.187)
Constant	9.754***	9.904***
	(0.431)	(0.442)
Observations	3142	3142
R^2	0.250	0.253
Adjusted R^2	0.241	0.243

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

In the OLS regression results, key factors significantly affect log revenue in Google Play store apps. Ratings, with a coefficient of 0.521, show a strong positive effect, suggesting higher-rated apps tend to earn more. Log Price, too, positively influences revenue but to a lesser degree (coefficient = 0.148). Monetization Strategy Number and Age Target Category both have notable positive impacts. Additionally, the Number of Languages and Category Number contribute to revenue generation, albeit to a smaller extent.

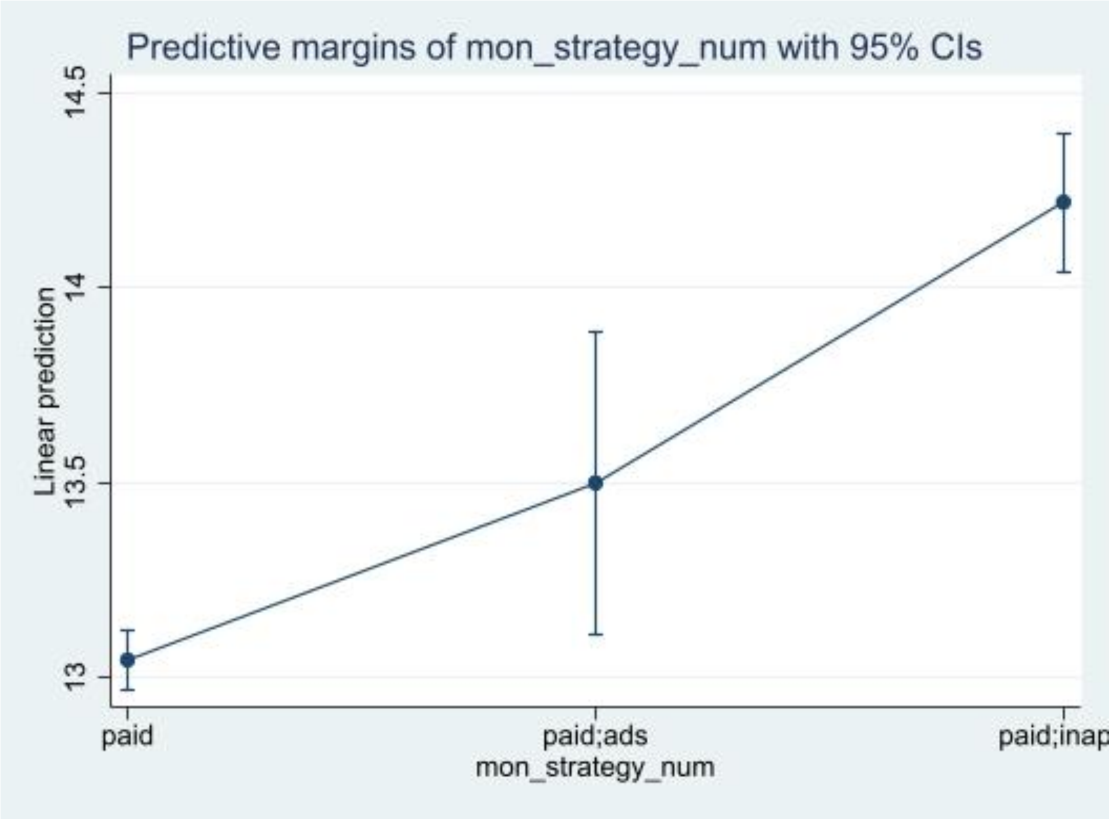

Figure 7: Linear prediction for mon_strategy_num

The graph shows the predicted margins with 95% confidence intervals for a linear model for the variable mon_strategy_num, which is plotted on the x-axis and is separated into three groups. The group labelled as "paid" displays the lowest estimate, exceeding 13, and a narrow 95% confidence interval indicating little uncertainty. The estimate for the "paid; ads" group is around 13.5, with a wider confidence interval that suggests greater uncertainty. With an estimate that is closest to 14.5 and a confidence interval that is the broadest, the "paid;inap" group has the highest level of uncertainty along with the strongest anticipated result. The pattern points to a potential rising association between the anticipated values and the categories of the variable.

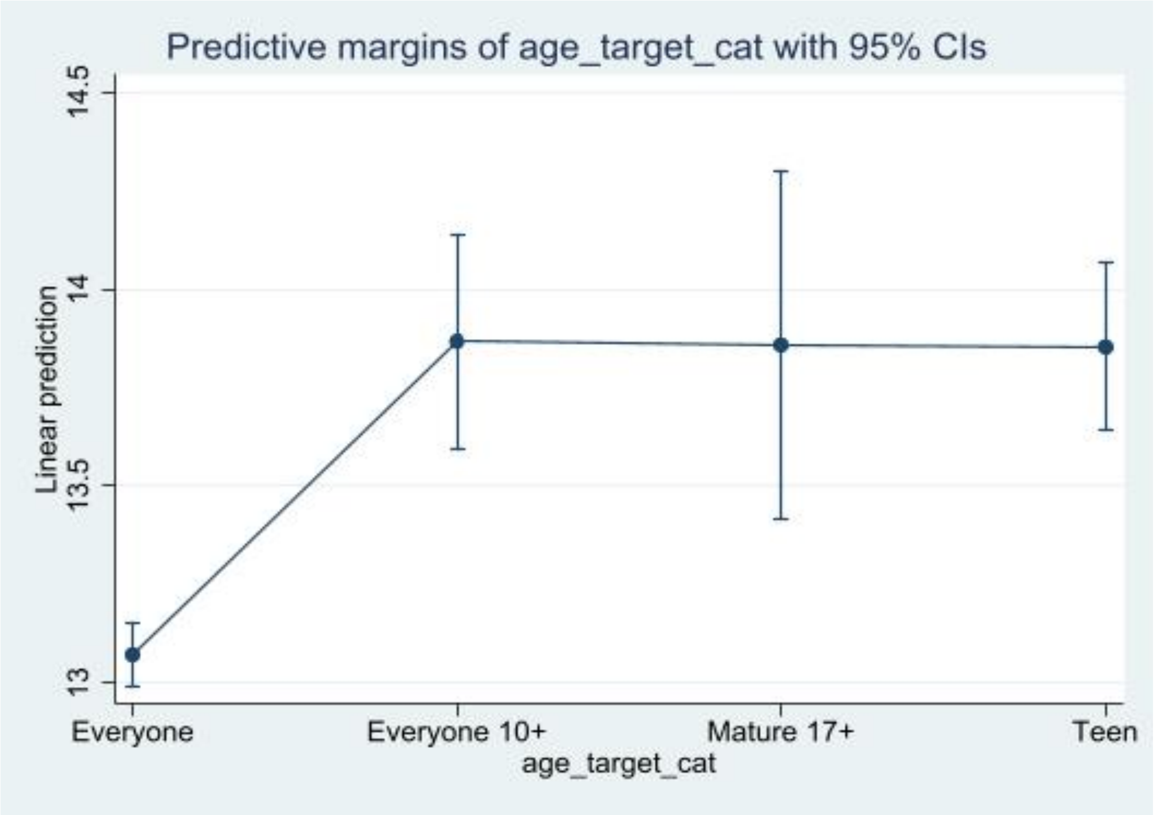

Figure 8: Predictive Margins for Different Age Categories

The graph shows that the estimated average effect for “Everyone" group is at the lower end of the range and has a wider confidence interval, indicating a higher level of uncertainty. In comparison to the "Everyone" group, the calculated mean effect for the "Everyone 10+" group is significantly higher and has a narrower confidence interval. Compared to the baseline "Everyone" group, this implies a stronger average effect with a higher degree of certainty in its measurement. The projected impacts for the "Teen" and "Mature 17+" categories are fairly like each other and closely match the "Everyone 10+" category. The overlapping confidence intervals between these categories, which imply that there aren't any appreciable variations in the expected values, provide evidence of this.

Implementing OLS regression with interaction terms. (Refer Table 6)

In the OLS regression model including interaction terms, significant main effects of rating, log_price, age_target_cat, num_langs, and category_num are seen. The introduction of interaction terms, particularly between 'paid;ads' and rating, shows a significant negative effect, showing that the impact of ratings on log revenue changes based on monetization strategy. This model, with an adjusted R-squared of 0.2529, shows enhanced explanatory power compared to the earlier version without interaction terms.

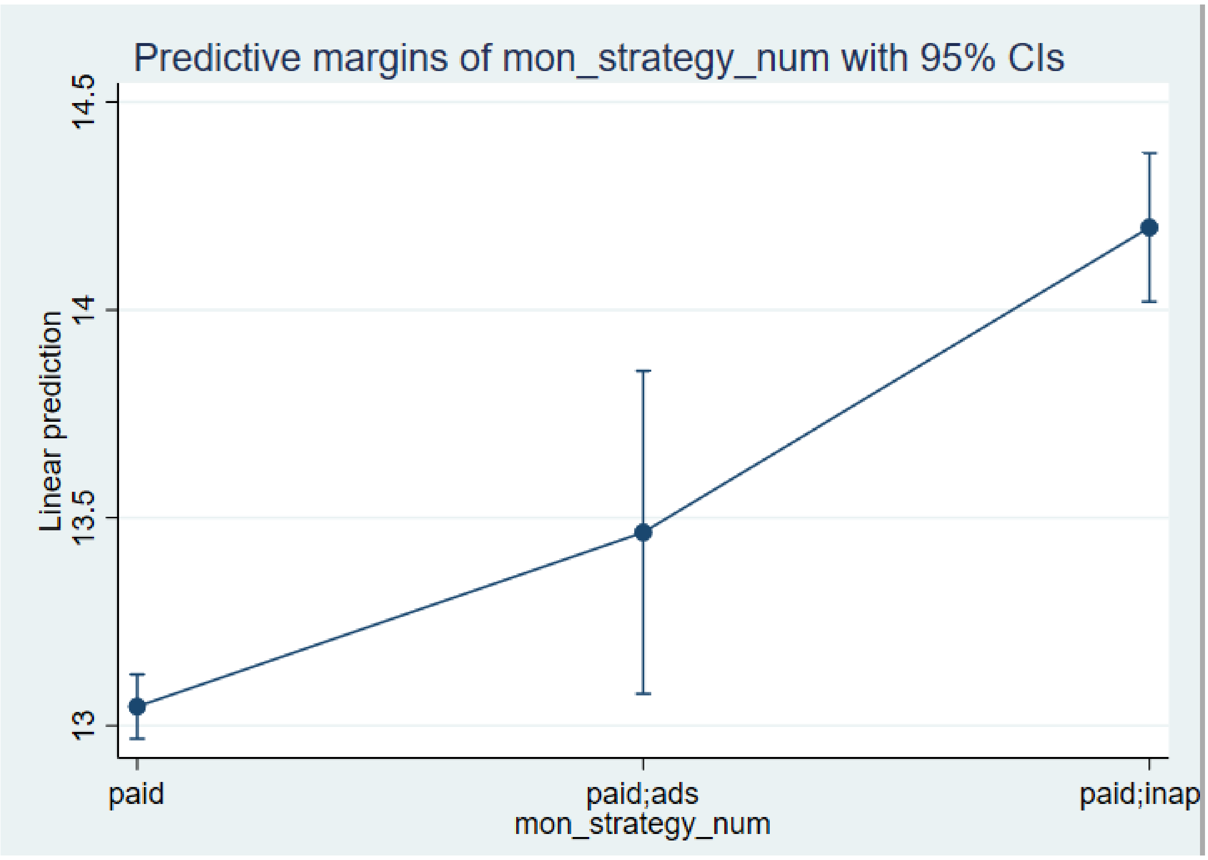

Figure 9: Prediction margin for mon_strategy_num with 95% cls

The graph depicts expected results depending on various monetization techniques, as indicated by the variable "mon_strategy_num." The linear forecasts show a clear increasing trend as the monetization approach grows more complex, beginning with a simple "paid" model, progressing to a "paid advertising" model, and finally concluding with a "paid in-app purchases" model.

According to the graph, there is a positive relationship between the details of the monetization approach and the expected outcome. This could mean that when businesses implement more complex monetization tactics, their projected returns, engagement, or other target metrics may rise. The confidence intervals reveal that forecasts for the simpler "paid" and "paid;ads" strategies are produced with greater precision, whereas the "paid;inap" approach, while exhibiting the highest precision, also has the lowest precision.

Robustness Analysis:

Table 7: (Model 3) Baseline model for robustness check with standard errors & (Model 4) Quadratic term for price to the model.

	(Model 3)	(Model 4)
rating	0.521***	0.517***
	(0.067)	(0.061)
log_price	0.149**	0.594***
	(0.047)	(0.107)
paid	0.000	0.000
	(.)	(.)
paid;ads	0.455*	0.455*
	(0.221)	(0.201)
paid;inapp	1.173***	1.182***
	(0.098)	(0.099)
Everyone	0.000	0.000
	(.)	(.)
Everyone 10+	0.795***	0.822***
	(0.153)	(0.148)
Mature 17+	0.787**	0.821***
	(0.299)	(0.230)
Teen	0.782***	0.824***
	(0.127)	(0.119)
num_langs	0.024***	0.024***
	(0.002)	(0.002)

Art & Design	Included	Included
log_price_sq		-0.157^{***} (0.034)
Constant	9.754^{***} (0.378)	9.545^{***} (0.432)
Observations	3142	3142
R^2	0.250	0.255
Adjusted R^2	0.241	0.246

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The baseline model with robust standard errors supplies more reliable estimates in the presence of heteroskedasticity. The results stay statistically significant for all variables, showing robustness. The coefficients of 'rating', 'log_price', 'mon_strategy_num', 'age_target_cat', 'num_langs', and 'category_num' continue to show significant impacts on log revenue. The use of robust standard errors adjusts for the heteroskedasticity detected earlier, ensuring that the statistical inferences made from the model are more correct and dependable.

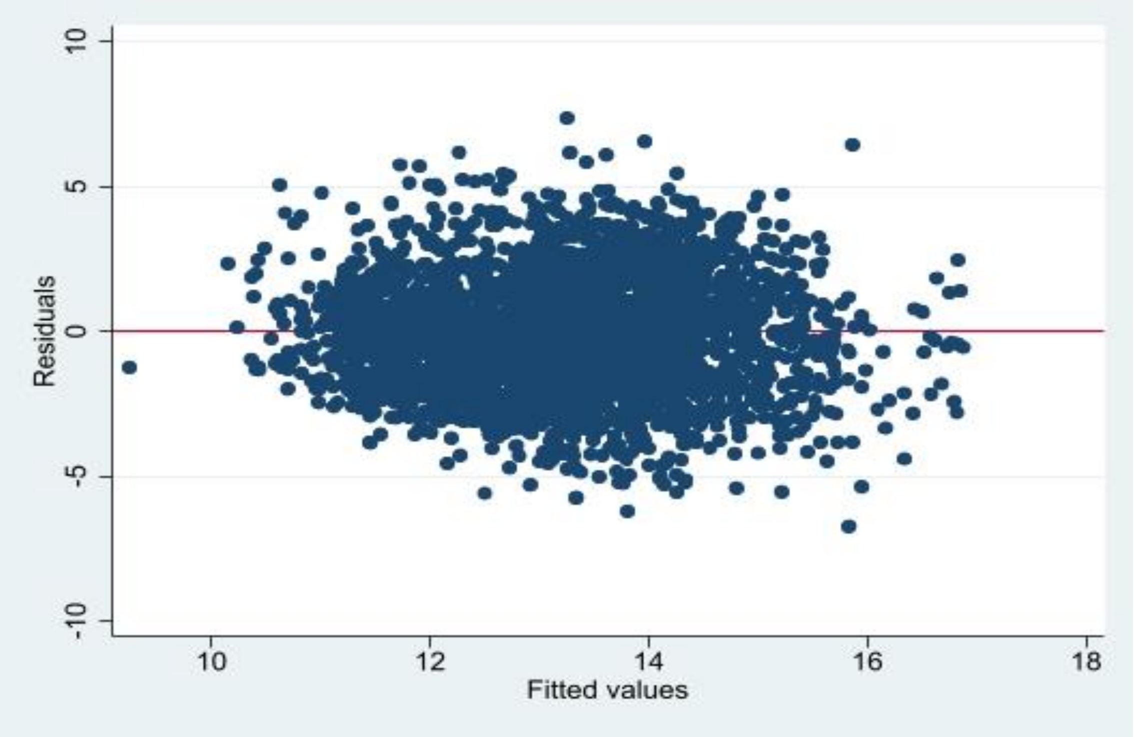

Figure 10: Scatter plot for residual and fitted values.

The given plot, a residual vs fitted plot, is frequently used to evaluate how well a linear regression model works. The dispersion of points to a random residual dispersion, indicating a model free from systematic mistakes that reflect the underlying relationship. The model's predictions are often neither too high nor too low, as indicated by the residuals, which primarily cluster around the horizontal line at zero. In general, the plot indicates that the model is operating as expected.

Heteroskedasticity Evaluation:

Table 8: Testing for heteroskedasticity using Breusch-Pagan test

White's test			
H0: Homoskedasticity			
Ha: Unrestricted heteroskedasticity			
chi2(227) = 340.94			
Prob > chi2 = 0.0000			
Cameron & Trivedi's decomposition of IM-test			
Source	chi2	df	p
Heteroskedasticity	340.94	227	0
Skewness	52.82	38	0.0556
Kurtosis	0.05	1	0.8319
Total	393.81	266	0

The Breusch–Pagan/Cook–Weisberg and White's tests in the regression model reveal heteroskedasticity. The Breusch–Pagan test, with a chi-square of 340.94 and p-value 0.0000, and White's test, with a chi-square of 340.94 and p-value 0.0000, both reject the null hypothesis of constant variance. This suggests varying residual variances across distinct levels of independent variables, potentially affecting the precision of coefficient estimates.

Regression - Quadratic Term. (Refer Table 7)

Incorporating a quadratic term for log_price reveals a significant non-linear impact on log_revenue: a positive coefficient for log_price (0.5944) and a negative for its squared term (-0.1566). This implies initial revenue increase with rising prices, then a potential decrease.

The model's adjusted R-squared increases to 0.2459, supporting the significance of other variables like rating and monetization strategy.

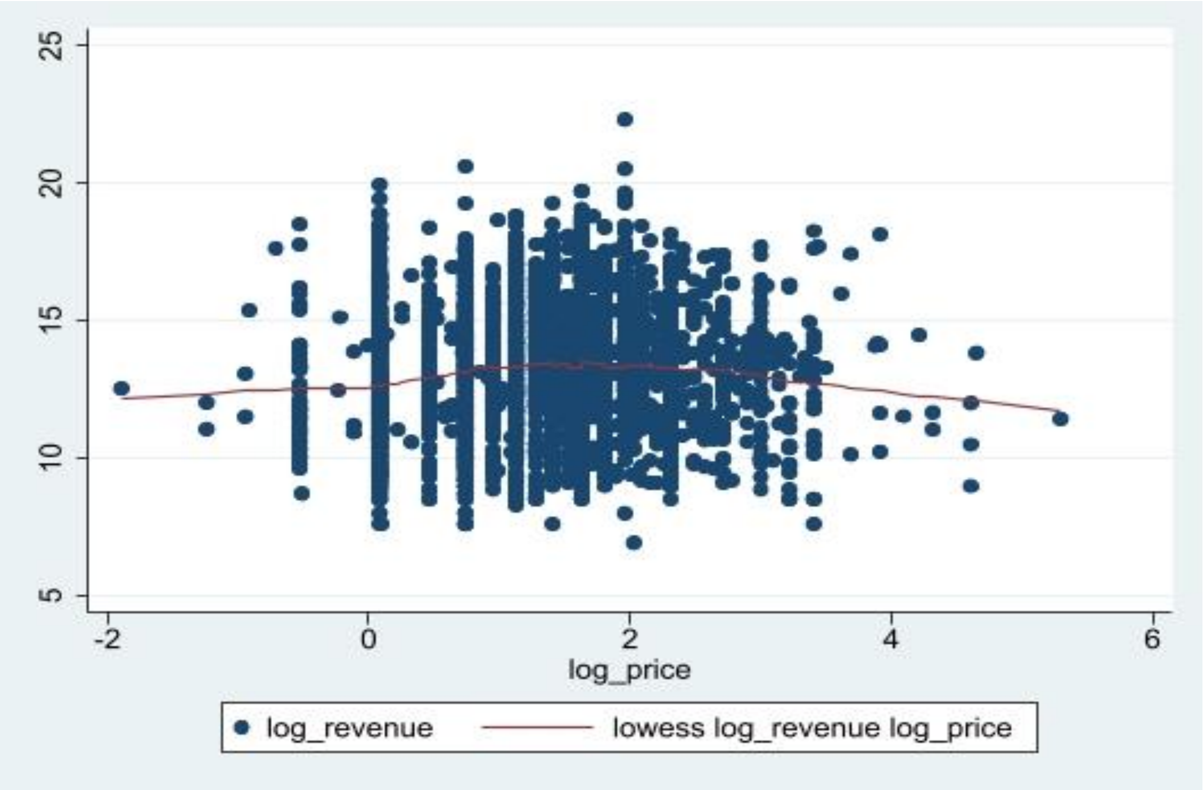

Figure 11: Logged Revenue vs Logged Price

The scatter plot titled "Logged Revenue vs. Logged Price" illustrates the relationship between two financial metrics in the app market. Data points standing for individual apps show how revenue (y-axis) changes with price (x-axis). The upward trend suggests a positive correlation:

as the price increases, revenue often does too. However, the pattern is not strictly linear, showing complexity in how price influences revenue. There's also noticeable variability, implying other factors are at play. Some data points fall outside the main cluster, potentially signalling outliers that may require further investigation.

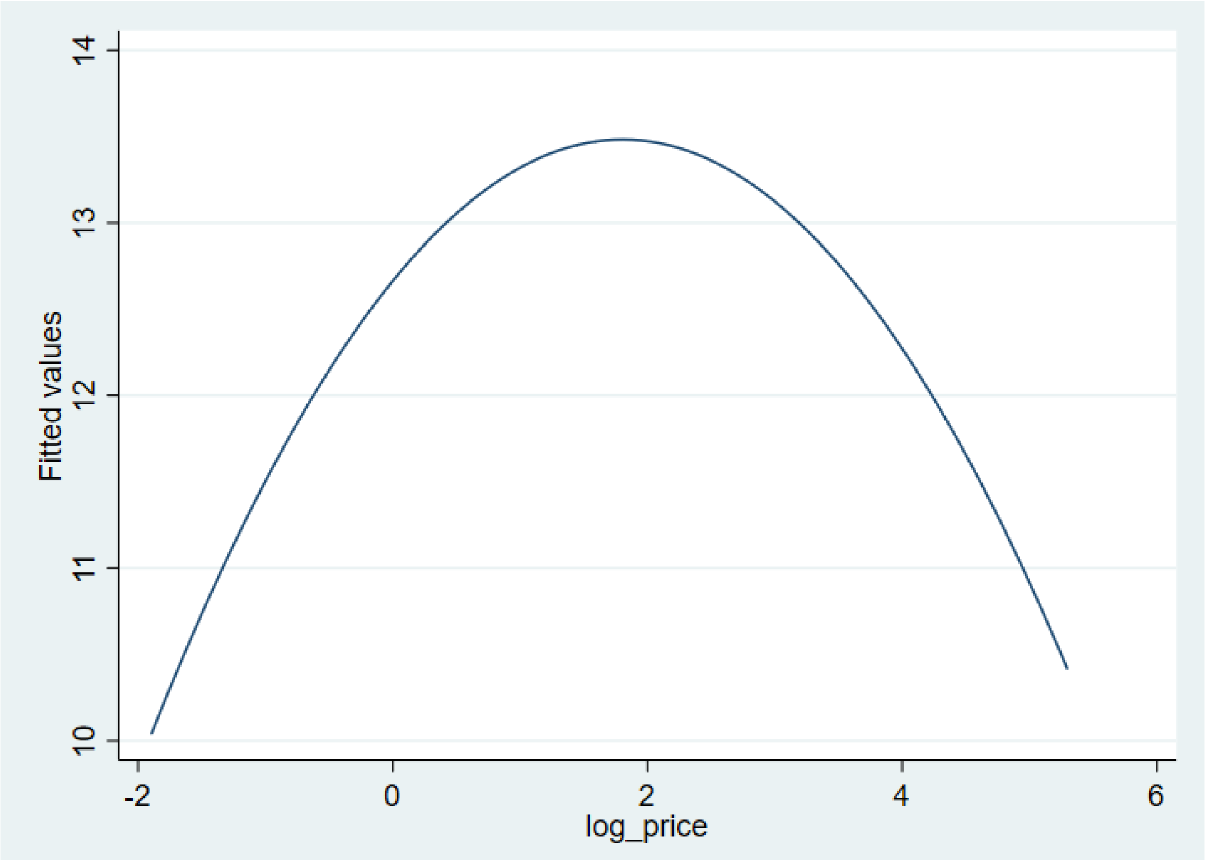

Figure 12: Predicted Log Revenue vs Log Price

The graph shows an arch-shaped, smooth trend connecting two metrics. The expected metric initially rises and reaches a maximum value as the price increases (after logarithmic adjustment). Nevertheless, after reaching this peak, an added rise in the logarithmic price is associated with a decline in the expected measure. This suggests that the highest expected result can be achieved at a best price level. The consistent representation of the trend in the dataset by the underlying model is shown by the homogeneity of the curve.

Endogeneity :

In the baseline regression model for app revenue, several endogeneity issues could skew the results. Not accounting for factors such as app quality, developer reputation, or marketing intensity—if they're correlated with variables like app ratings—could introduce bias. Additionally, if revenue or ratings are inaccurately measured, this would distort the analysis. There's also a possibility of simultaneity; revenue may influence ratings through increased investment in app quality. Furthermore, if app prices are set based on unobserved quality expectations, the price variable itself may be endogenous.

To refine the model, variables like user engagement, frequency of updates, and historical performance could be included to give a fuller picture and possibly correct for omitted variable bias. If panel data were available, it could further strengthen the model by allowing for fixed or random effects models that adjust for unobserved, consistent app characteristics. This would also open the use of instrumental variable regression to better address simultaneity and endogeneity, thereby improving causal inference.

Conclusion:

Primary conclusions are as follows:

App Features and Income: Important factors that affect app revenue include target age group, monetization techniques, user ratings, app price, and the number of languages offered. Apps with higher ratings usually bring in more money, which highlights how crucial user pleasure is. Furthermore, apps with broader language support typically generate more money, showing the advantage of a worldwide market strategy. Gaming vs. Non-Gaming applications according to the research, gaming applications often generate more money than non-gaming apps. Statistical analyses, including ANOVA and two-sample t-tests, which show substantial disparities in revenue creation between these groups, support this gap. Various monetization techniques, such as in-app purchases and adverts, have varying effects on income. Revenue increases up to a specific price point and then may decrease, indicating a non-linear relationship between price and revenue.

Model Robustness and Statistical Analysis of data is analysed in the report using a variety of statistical models and tests, such as Ordinary Least Square (OLS) regression. Tests for heteroskedasticity and residual analysis verify the robustness of these models. Regression models with interaction terms show the complex connections between various factors and revenue. Revenue Generation and Price Effects shows strong non-linear impact on revenue is indicated by the inclusion of a quadratic term for price in the regression model. According to the analysis, revenue initially rises when prices rise, but at a certain point, revenue may decline as prices rise. Endogeneity Considerations section of the analysis addresses possible endogeneity problems in the regression model, including the presence of unknown variables like marketing intensity and app quality. Panel data and the addition of further variables are recommended as ways to improve the model.

In conclusion, the study provides insightful information on the variables affecting app sales on the Google Play store, highlighting the significance of monetization techniques, smart pricing, and user pleasure. It also identifies areas that should be investigated in the future to address possible biases and improve the explanatory power of the model.