	Rating	Reviews	Size	Installs	Price	Android Ver
Rating	1.000	0.069	0.079	0.051	-0.022	0.063
Reviews	0.069	1.000	0.120	0.633	-0.010	0.033
Size	0.079	0.120	1.000	0.057	-0.026	0.124
Installs	0.051	0.633	0.057	1.000	-0.011	0.045
Price	-0.022	-0.010	-0.026	-0.011	1.000	0.006
Android Ver	0.063	0.033	0.124	0.045	0.006	1.000

Figure 1: Correlation matrix output, showing comparison results across all dataset categories

Figure 2: WordCloud visualization of the 500 most common words in the App Titles of our dataset.

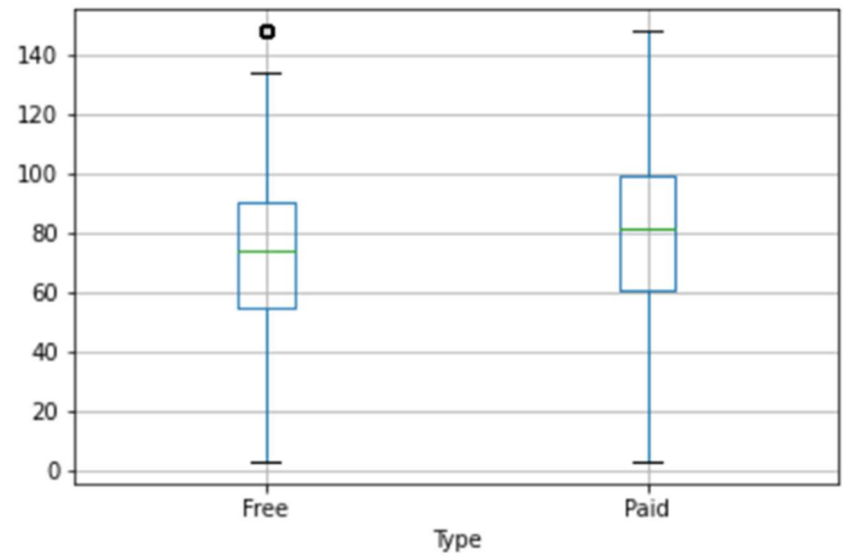

Figure 3: Boxplot Free VS Paid

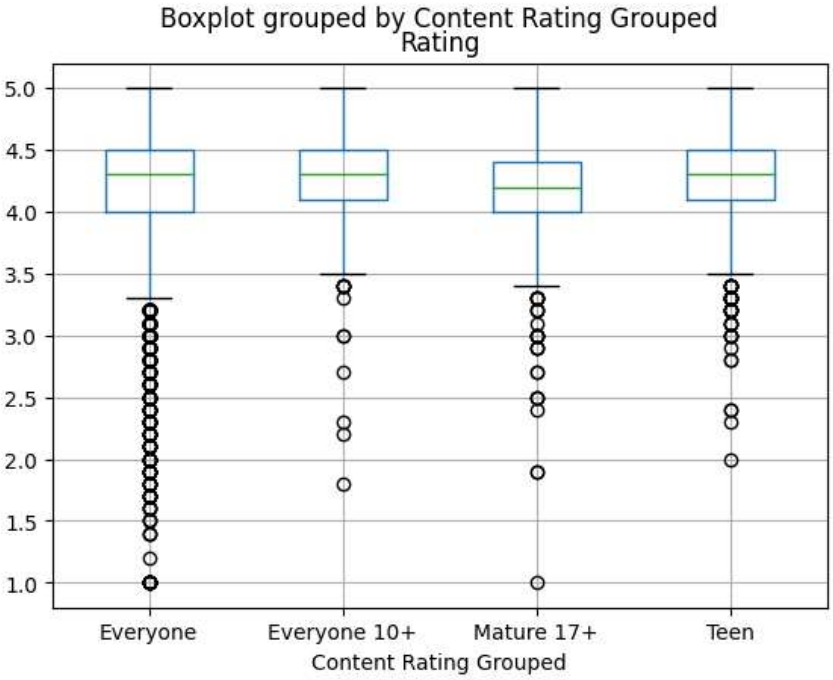

Figure 4: Boxplot grouped by Content Rating Grouped Rating

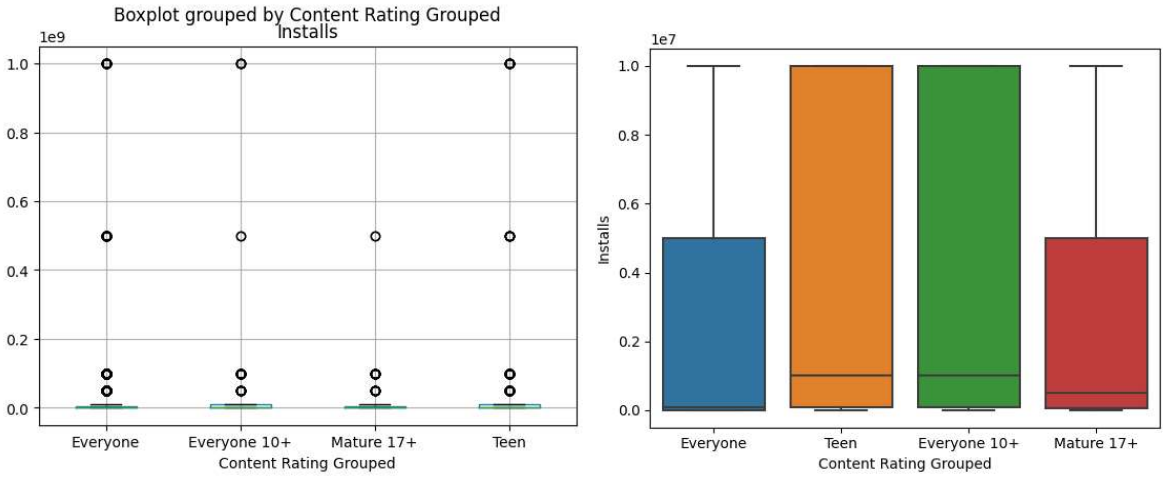

Figure 5 (a) & (b): Boxplot grouped by Content Rating Grouped with outliers and with outliers (left to right)

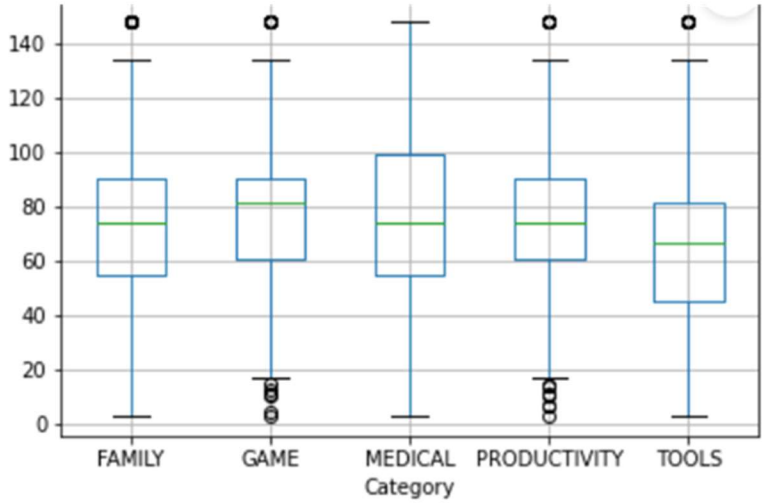

Figure 6: Boxplot of top five app categories.

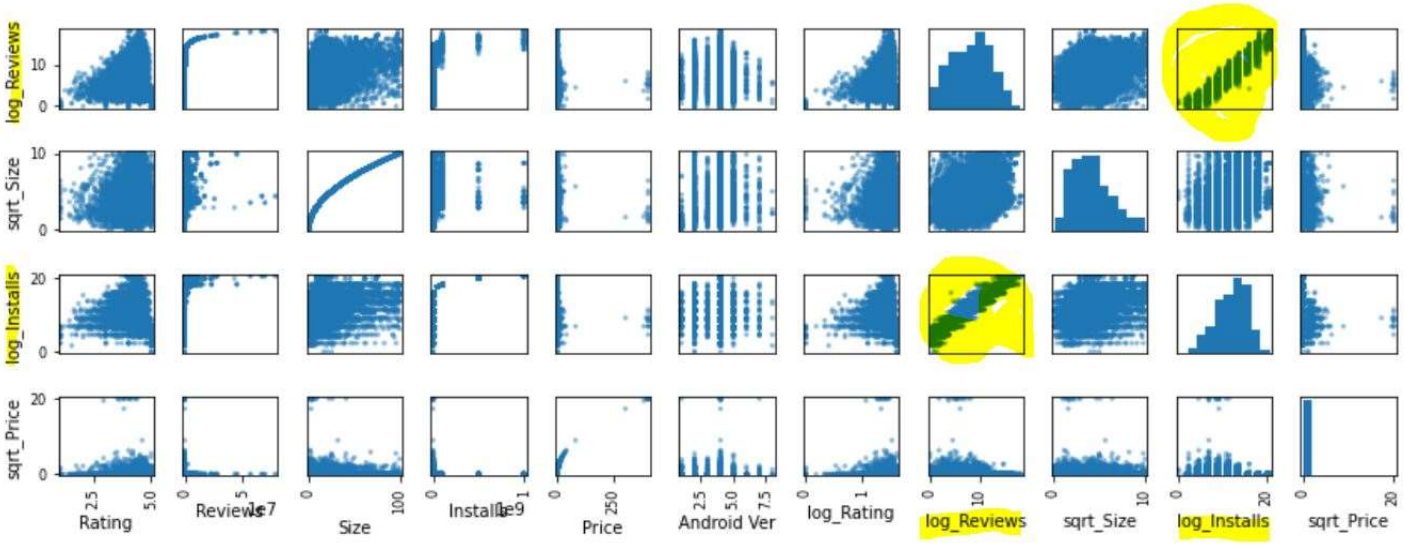

Figure 7: scatter matrix plot after transformation

Figure 8. Scatter matrix plot for numerical variables

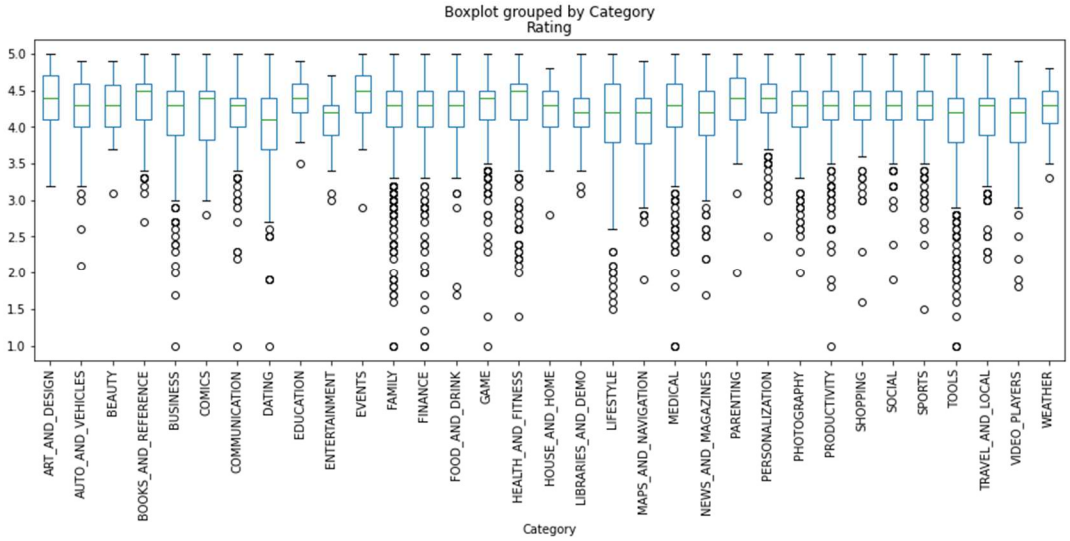

Figure 9. Boxplot of Rating grouped by Category.

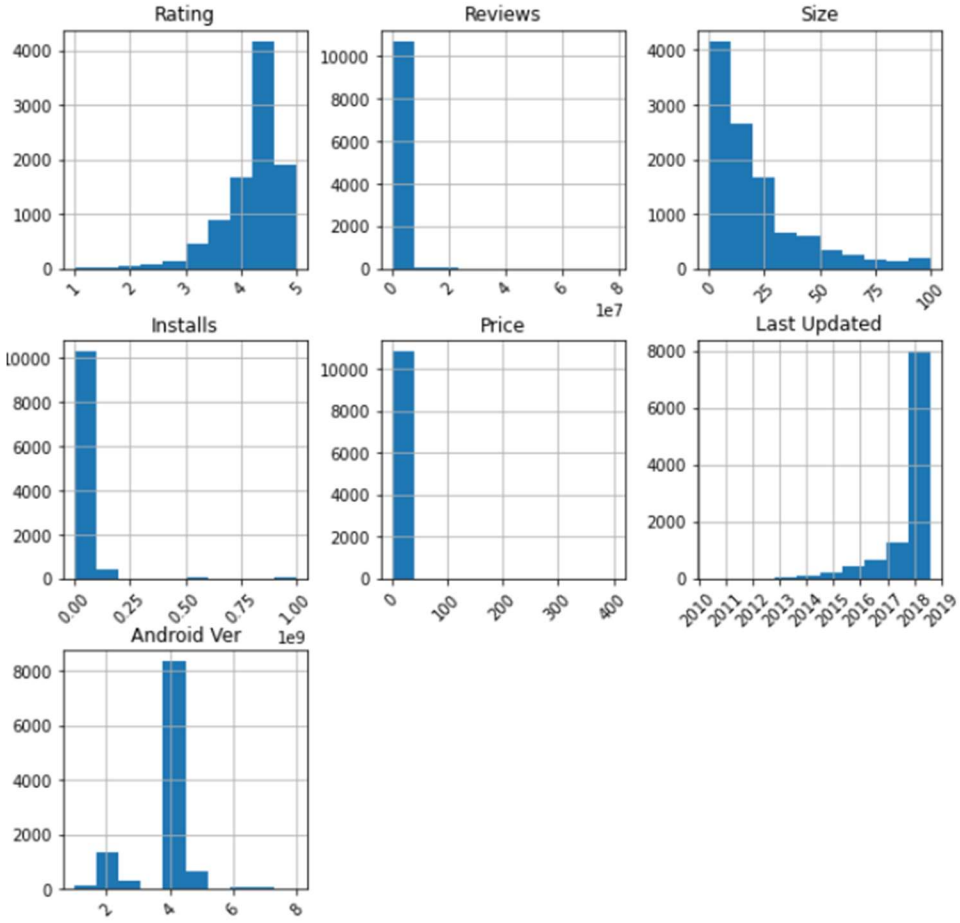

Figure 10. Histograms of numerical variables.

```
# Fit the model
df_nonan=df.dropna()
df_nonan['Typen'] = df_nonan['Type'].replace({'Free':'0','Paid':'1'}, regex=True).astype(int)
predictors = ['Reviews', 'Price', 'Size']
m = Logit(df_nonan['Typen'], df_nonan[predictors])
m = m.fit()
```

```
Warning: Maximum number of iterations has been exceeded.
         Current function value: inf
         Iterations: 35
```

Figure 11. Logistic Regression failure.

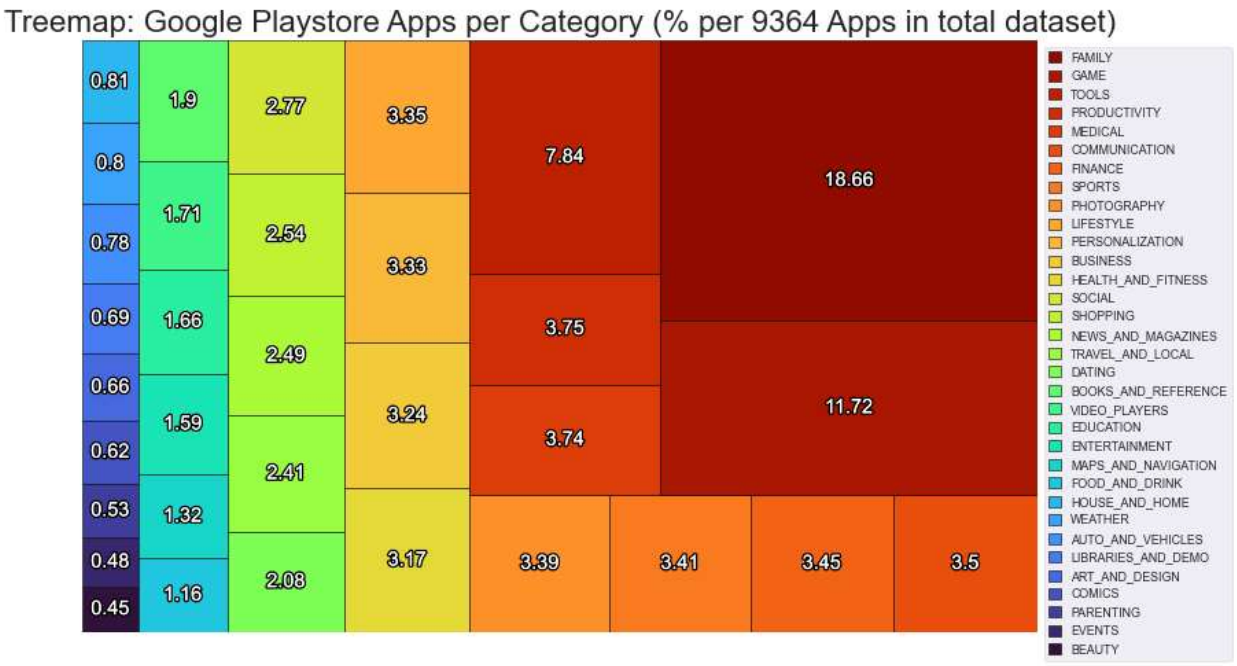

Figure 12: Proportional Treemap of App Categories in Google Play Store Kaggle dataset, by % of total.

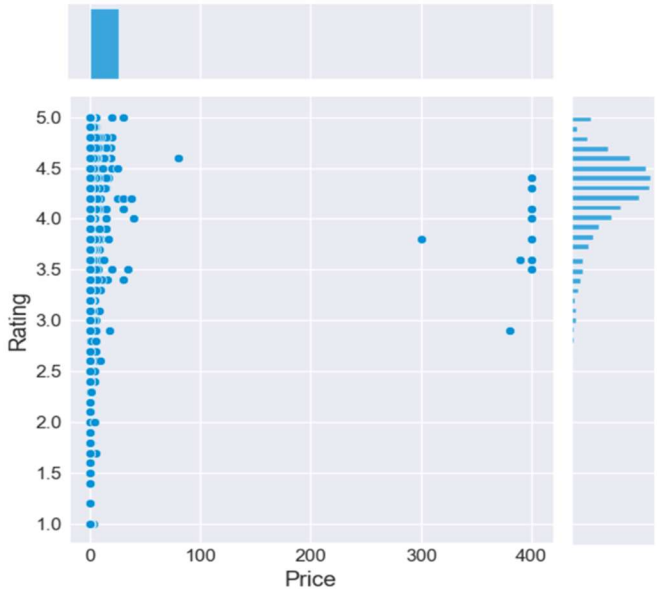

Figure 13: Seaborn jointplot (scatterplot + histogram) of Rating vs. Price for Paid apps.

Table 3: Filtered list of most expensive apps (>$250-$400) in the dataset

	Category	App
2876	FAMILY	most expensive app (H)
3004	LIFESTYLE	💎 I'm rich
3007	LIFESTYLE	I'm Rich - Trump Edition
3722	LIFESTYLE	I am rich
3725	FAMILY	I am Rich Plus
3726	LIFESTYLE	I am rich VIP
3727	FINANCE	I Am Rich Premium
3728	LIFESTYLE	I am extremely Rich
3729	FINANCE	I am Rich!
3732	FAMILY	I Am Rich Pro
3734	FINANCE	I am rich (Most expensive app)
3736	FAMILY	I Am Rich
3739	FINANCE	I am Rich
3743	FINANCE	I AM RICH PRO PLUS
7541	FINANCE	I am rich(premium)

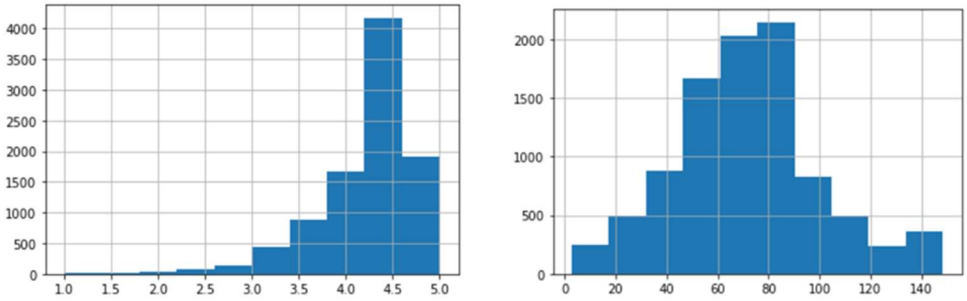

Figure 14: Histogram of Rating before and after transformation.

group1	group2	stat	pval	pval_corr	reject
Adults only 18+	Everyone	-0.2619	0.7934	1.0	False
Adults only 18+	Everyone 10+	-0.478	0.6329	1.0	False
Adults only 18+	Mature 17+	-0.4622	0.6442	1.0	False
Adults only 18+	Teen	-0.3061	0.7596	1.0	False
Adults only 18+	Unrated	nan	nan	nan	False
Everyone	Everyone 10+	-9.8363	0.0	0.0	True
Everyone	Mature 17+	-0.2	0.8415	1.0	False
Everyone	Teen	-6.8558	0.0	0.0	True
Everyone	Unrated	nan	nan	nan	False
Everyone 10+	Mature 17+	4.3003	0.0	0.0003	True
Everyone 10+	Teen	1.9607	0.0501	0.7517	False
Everyone 10+	Unrated	nan	nan	nan	False
Mature 17+	Teen	-2.2718	0.0232	0.3487	False
Mature 17+	Unrated	nan	nan	nan	False
Teen	Unrated	nan	nan	nan	False

Figure 15: Rating by Content Rating ANOVA comparison table.

Test Multiple Comparison ttest_ind FWER=0.05 method=bonf
alphacSidak=0.01, alphacBonf=0.005

group1	group2	stat	pval	pval_corr	reject
FAMILY	GAME	-3.2366	0.0012	0.0122	True
FAMILY	MEDICAL	-2.2072	0.0274	0.2741	False
FAMILY	PRODUCTIVITY	-0.4319	0.6659	1.0	False
FAMILY	TOOLS	5.6146	0.0	0.0	True
GAME	MEDICAL	-0.3712	0.7106	1.0	False
GAME	PRODUCTIVITY	1.8248	0.0682	0.6823	False
GAME	TOOLS	8.7441	0.0	0.0	True
MEDICAL	PRODUCTIVITY	1.3518	0.1769	1.0	False
MEDICAL	TOOLS	5.4464	0.0	0.0	True
PRODUCTIVITY	TOOLS	4.3019	0.0	0.0002	True

Figure 16: Rating by Category ANOVA comparison table

	Rating	Reviews	Size	Installs	Price	Android Ver	log_Rating	log_Reviews	sqrt_Size	log_Installs	sqrt_Price
Rating	1.000000	0.068753	0.078949	0.050909	-0.022353	0.063155	0.979145	0.206023	0.087028	0.114569	-0.002459
Reviews	0.068753	1.000000	0.119601	0.633426	-0.009559	0.033414	0.064446	0.314715	0.116646	0.280300	-0.026445
Size	0.078949	0.119601	1.000000	0.056549	-0.025503	0.124292	0.080482	0.313536	0.965735	0.274296	-0.029456
Installs	0.050909	0.633426	0.056549	1.000000	-0.011330	0.045001	0.051860	0.331783	0.070131	0.346813	-0.031801
Price	-0.022353	-0.009559	-0.025503	-0.011330	1.000000	0.005685	-0.017719	-0.042510	-0.032359	-0.059392	0.883153
Android Ver	0.063155	0.033414	0.124292	0.045001	0.005685	1.000000	0.052244	0.089601	0.182130	0.095235	-0.038571
log_Rating	0.979145	0.064446	0.080482	0.051860	-0.017719	0.052244	1.000000	0.227306	0.088247	0.148863	-0.002416
log_Reviews	0.206023	0.314715	0.313536	0.331783	-0.042510	0.089601	0.227306	1.000000	0.357095	0.957461	-0.112234
sqrt_Size	0.087028	0.116646	0.965735	0.070131	-0.032359	0.182130	0.088247	0.357095	1.000000	0.318724	-0.042272
log_Installs	0.114569	0.280300	0.274296	0.346813	-0.059392	0.095235	0.148863	0.957461	0.318724	1.000000	-0.166709
sqrt_Price	-0.002459	-0.026445	-0.029456	-0.031801	0.883153	-0.038571	-0.002416	-0.112234	-0.042272	-0.166709	1.000000

Figure 17: Correlation table after transformation.

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.917
Model:	OLS	**Adj. R-squared:**	0.917
Method:	Least Squares	**F-statistic:**	9.785e+04
Date:	Sat, 08 Apr 2023	**Prob (F-statistic):**	0.00
Time:	09:55:23	**Log-Likelihood:**	-13619.
No. Observations:	8890	**AIC:**	2.724e+04
Df Residuals:	8888	**BIC:**	2.726e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-3.5578	0.040	-90.008	0.000	-3.635	-3.480
x	0.9682	0.003	312.814	0.000	0.962	0.974

Omnibus:	72.090	**Durbin-Watson:**	1.691
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	107.807
Skew:	-0.067	**Prob(JB):**	3.89e-24
Kurtosis:	3.522	**Cond. No.**	42.7

Figure 18: Regression Model Summary.

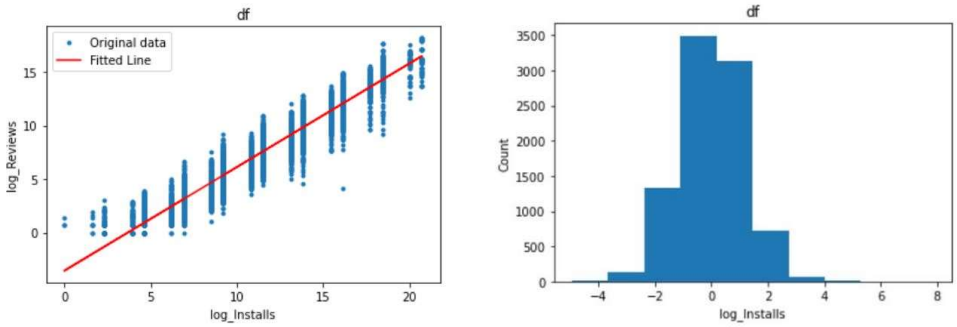

Figure 19. Residual plot of Regression Model.