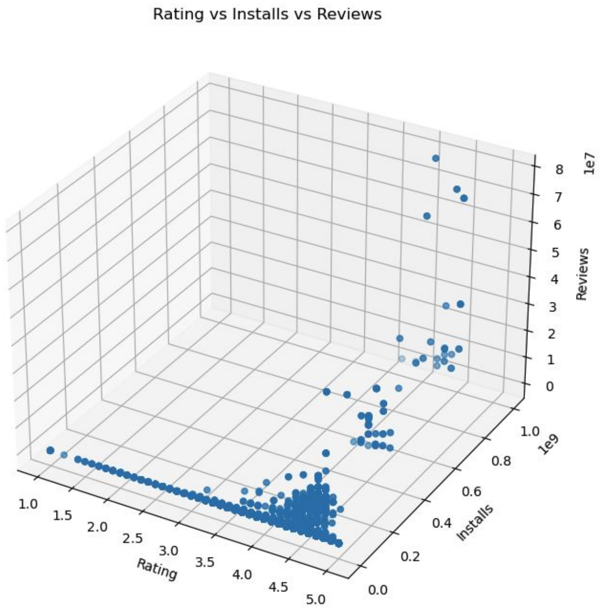
Does price of the application impact its quality?

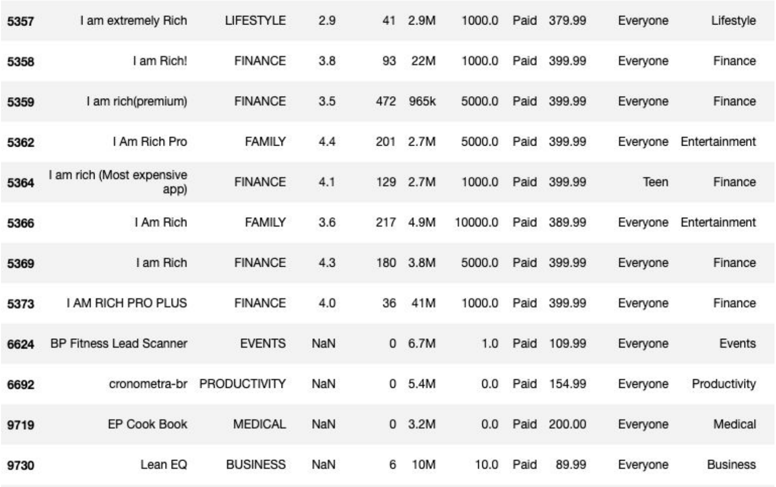
Michal Borek - CMSE 201 (Section 5)

Why did I do it?

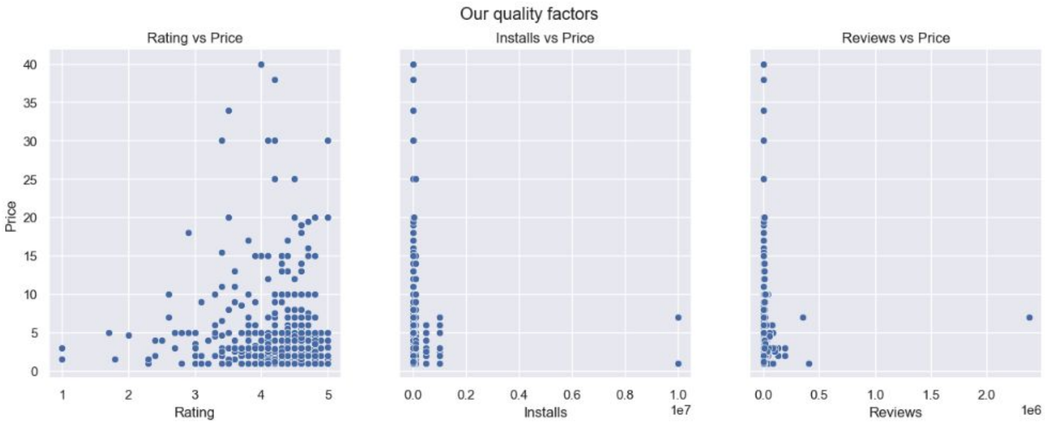
Kaggle:
https://www.kaggle.com/datasets/lava18/google-play-store-apps

Cleaning Data


Rating vs Installs vs Reviews



5357	I am extremely Rich	LIFESTYLE	2.9	41	2.9M	1000.0	Paid	379.99	Everyone	Lifestyle
5358	I am Rich!	FINANCE	3.8	93	22M	1000.0	Paid	399.99	Everyone	Finance
5359	I am rich(premium)	FINANCE	3.5	472	965k	5000.0	Paid	399.99	Everyone	Finance
5362	I Am Rich Pro	FAMILY	4.4	201	2.7M	5000.0	Paid	399.99	Everyone	Entertainment
5364	I am rich (Most expensive app)	FINANCE	4.1	129	2.7M	1000.0	Paid	399.99	Teen	Finance
5366	I Am Rich	FAMILY	3.6	217	4.9M	10000.0	Paid	389.99	Everyone	Entertainment
5369	I am Rich	FINANCE	4.3	180	3.8M	5000.0	Paid	399.99	Everyone	Finance
5373	I AM RICH PRO PLUS	FINANCE	4.0	36	41M	1000.0	Paid	399.99	Everyone	Finance
6624	BP Fitness Lead Scanner	EVENTS	NaN	0	6.7M	1.0	Paid	109.99	Everyone	Events
6692	cronometra-br	PRODUCTIVITY	NaN	0	5.4M	0.0	Paid	154.99	Everyone	Productivity
9719	EP Cook Book	MEDICAL	NaN	0	3.2M	0.0	Paid	200.00	Everyone	Medical
9730	Lean EQ	BUSINESS	NaN	6	10M	10.0	Paid	89.99	Everyone	Business


Our quality factors

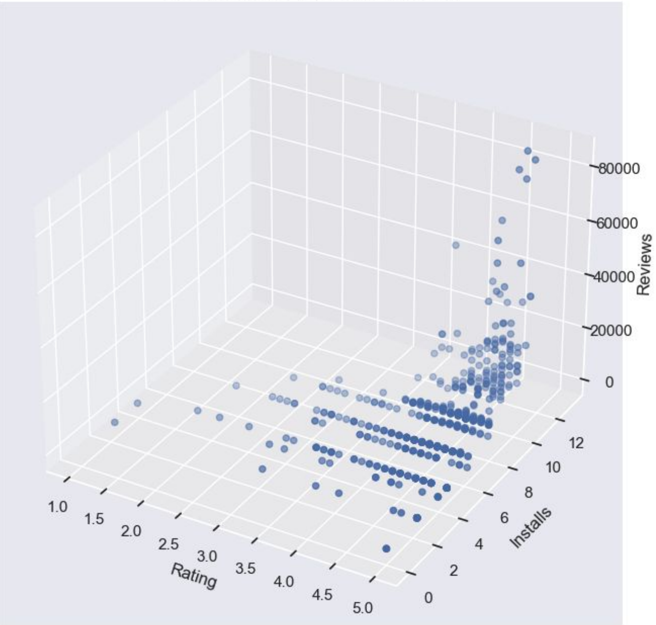
Rating vs Price

Installs vs Price

Reviews vs Price

Research

Rating vs Installs vs Reviews (Cleaned)

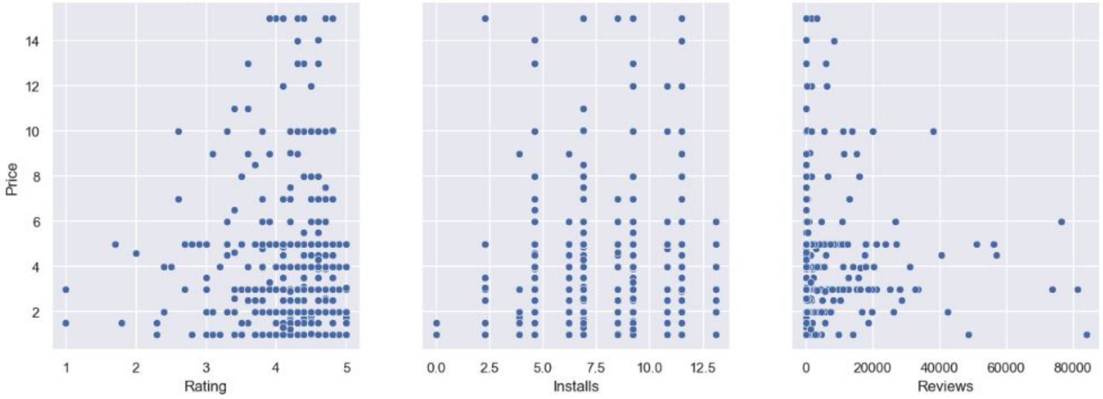


Our quality factors

Rating vs Price

Installs vs Price

Reviews vs Price

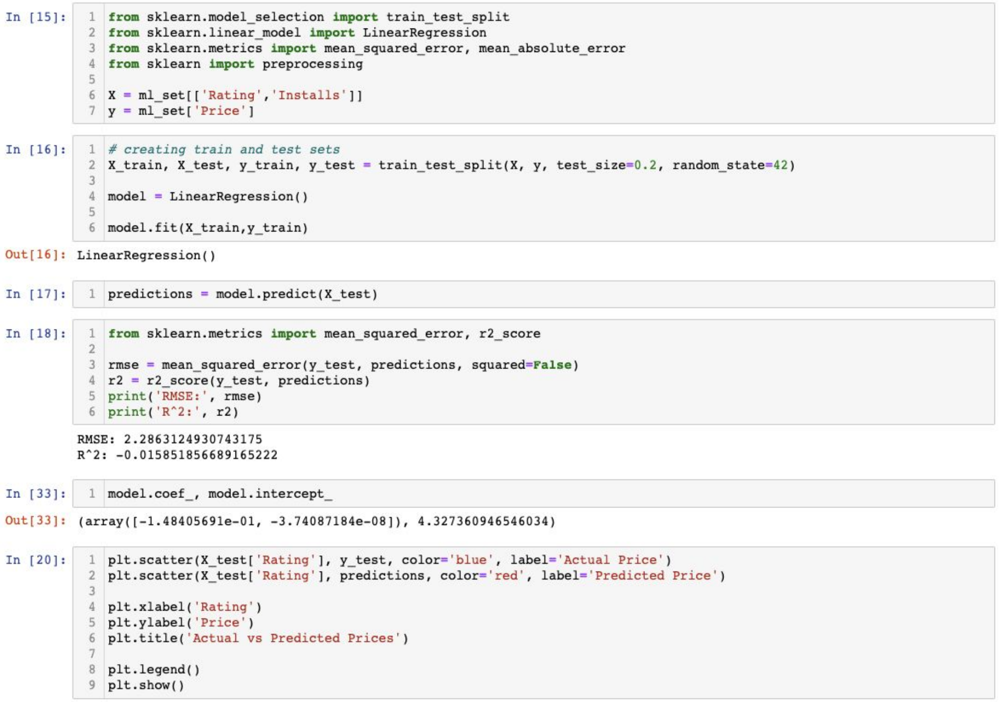


OUTLIERS

Apps with very high number of installs, such as apps with 10 million installs or more.
Apps with very low number of installs, such as apps with only 10 installs.
Apps with very low number of reviews, such as apps with only 1 or 2 reviews.
Apps with very low or very high ratings, such as apps with a rating of 1 or 5.

Introduction to Machine Learning

- Linear Regression
- RMSE
- R^2
- Coefficients and Intercept



```
In [15]:
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, mean_absolute_error
from sklearn import preprocessing

X = ml_set[['Rating','Installs']]
y = ml_set['Price']
```

```
In [16]:
# creating train and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

model = LinearRegression()

model.fit(X_train,y_train)
```

```
Out[16]: LinearRegression()
```

```
In [17]:
predictions = model.predict(X_test)
```

```
In [18]:
from sklearn.metrics import mean_squared_error, r2_score

rmse = mean_squared_error(y_test, predictions, squared=False)
r2 = r2_score(y_test, predictions)
print('RMSE:', rmse)
print('R^2:', r2)
```

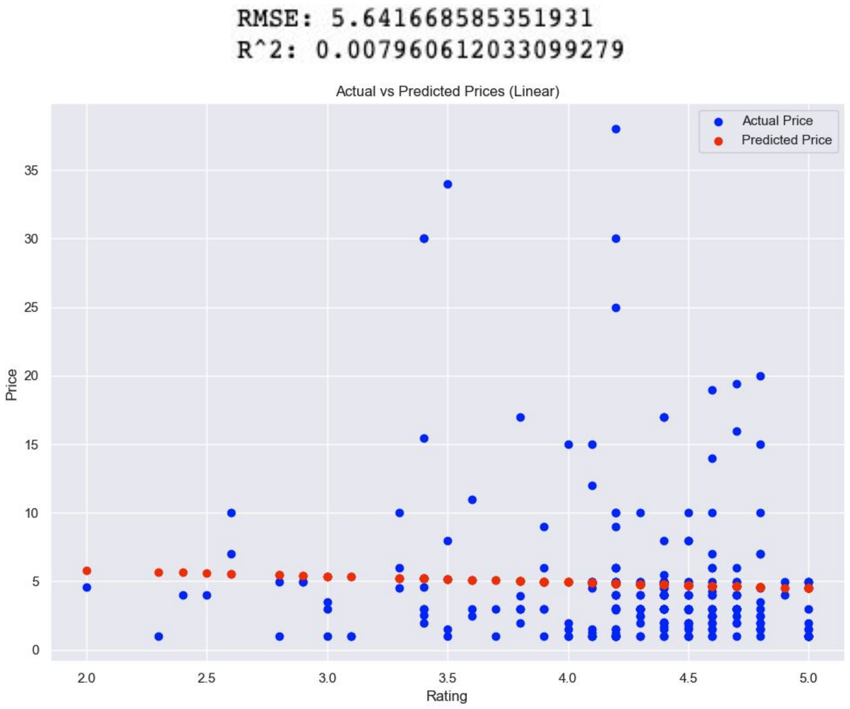
```
RMSE: 2.2863124930743175
R^2: -0.015851856689165222
```

```
In [33]:
model.coef_, model.intercept_
```

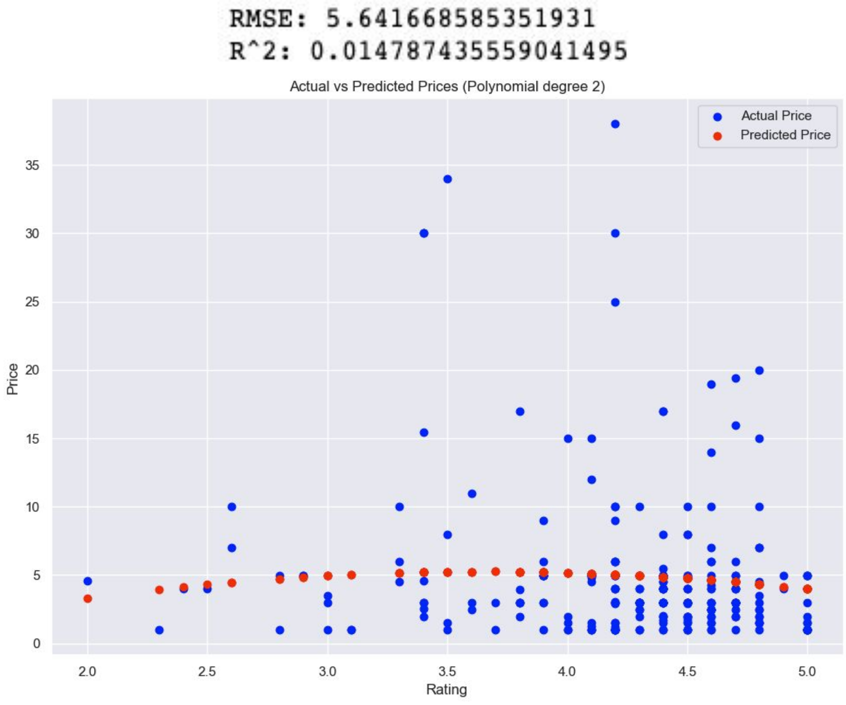
```
Out[33]: (array([-1.48405691e-01, -3.74087184e-08]), 4.327360946546034)
```

```
In [20]:
plt.scatter(X_test['Rating'], y_test, color='blue', label='Actual Price')
plt.scatter(X_test['Rating'], predictions, color='red', label='Predicted Price')

plt.xlabel('Rating')
plt.ylabel('Price')
plt.title('Actual vs Predicted Prices')

plt.legend()
plt.show()
```

Machine Learning Models



```
RMSE: 5.641668585351931
R^2: 0.007960612033099279
```



```
RMSE: 5.641668585351931
R^2: 0.014787435559041495
```

Conclusion

- We can conclude that there is small or no relationship between price and quality,
- Furthermore, we can deduce that developers prefer free apps with in-app purchases, and these apps may be more profitable than paid apps. However, we do not have enough data to establish the percentage of users that buy in-app products.