Features Leading to a Well-Received Video Game

by Brian Nguyen at Metis

TARGET AND SOURCE DATA

OVERALL USER RATING

Target metric (user rating) and features collected from store.steampowered.com.

Steam is a video game digital distribution service.

Exploratory approach to examine how features influence video game reception.

OVERVIEW OF DATA COLLECTED

32,537 Games (Initial Scrape)

2,871 Games (Clean/Undersample)

6 Features Examined

EDA: FEATURES AND SIGNIFICANCE

REVIEWS/DAY

All reviews since game release date normalized to reviews/day

RECENT REVIEWS

Number of recent reviews within 30 days (of July 10, 2020)

PRICE

Price of the game (discounts removed)

STORAGE

Minimum storage space required for installation

QTY GENRES

Quantity of genres that the game is listed under

INDIE

Categorization of independent game developer (1) or not (0)

EDA: FEATURES AND SIGNIFICANCE

Smallest Contribution to R^2

REVIEWS/DAY

All reviews since game release date normalized to reviews/day

RECENT REVIEWS

Number of recent reviews within 30 days (of July 10, 2020)

P-value = 0.12

PRICE

Price of the game (discounts removed)

Evaluate Features using Ordinary Least Squares (statsmodel)

STORAGE

Minimum storage space required for installation

QTY GENRES

Quantity of genres that the game is listed under

INDIE

Categorization of independent game developer (1) or not (0)

EDA: FEATURES AND SIGNIFICANCE

Smallest Contribution to R^2

REVIEWS/DAY

All reviews since game release date normalized to reviews/day

RECENT REVIEWS

Number of recent reviews within 30 days (of July 10, 2020)

P-value = 0.12

PRICE

Price of the game (discounts removed)

Total of <u>Four (4)</u> Features Will be Used for Modeling

Largest Contribution to R^2

STORAGE

Minimum storage space required for installation

QTY GENRES

Quantity of genres that the game is listed under

INDIE

Categorization of independent game developer (1) or not (0)

Multicollinearity

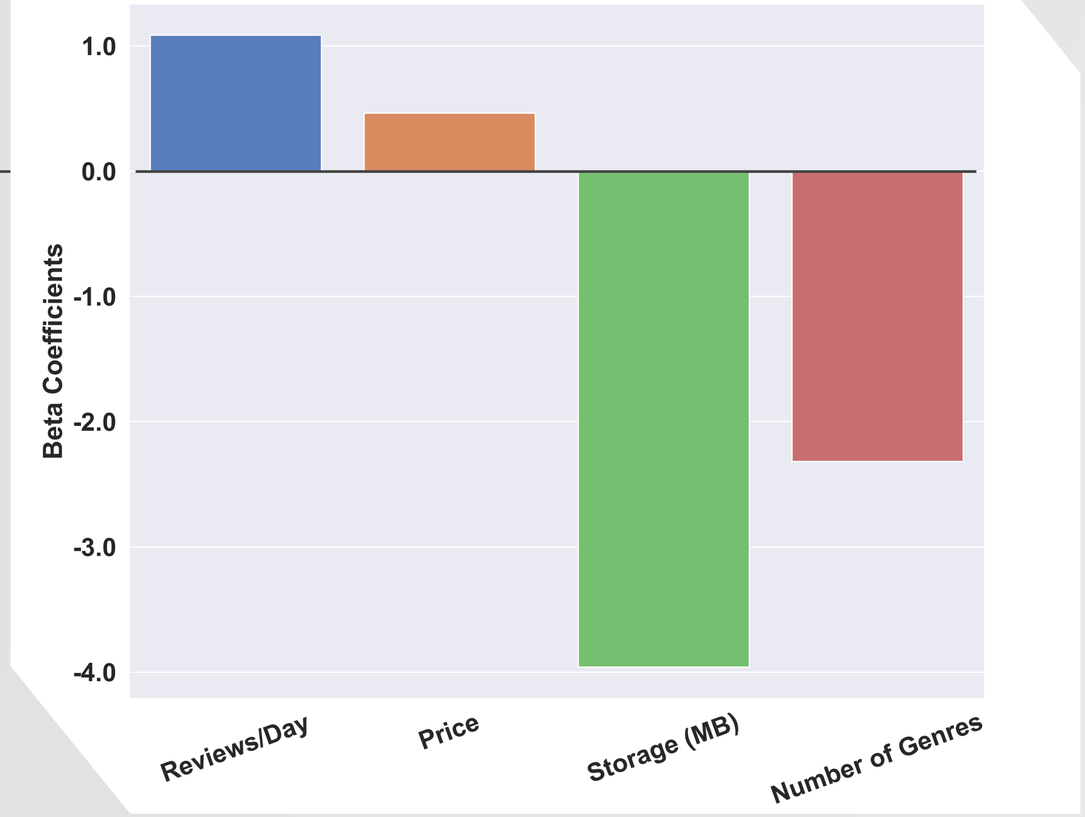
CROSS-VALIDATION AND TESTING

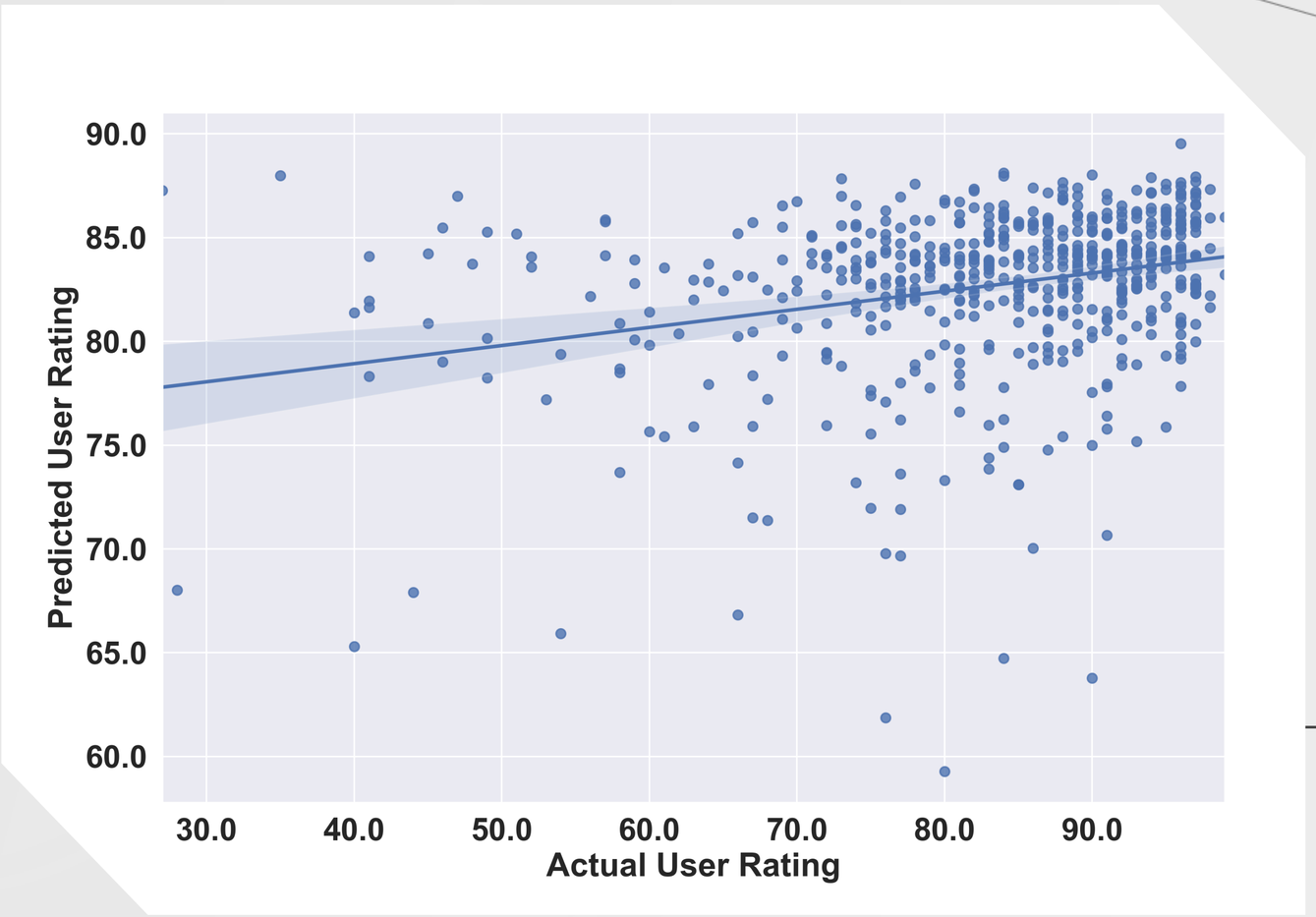
	$R^2_{val} \pm \sigma$	R^2_{test}	MAE
LINEAR	0.092 ± 0.016	0.073	9.1624
RIDGE (λ = 0.03)	0.092 ± 0.016	0.073	9.1624
LASSO (λ = 0.01)	0.092 ± 0.016	0.073	9.1618

R^2_{val}: validation with 20% data; K-fold = 5

R^2_{test}: testing with 20% data

β-COEFFICIENTS AND PREDICTING





BUSINESS TAKEAWAYS

(-4) STORAGE

Large storage games from AAA/large companies

AAA rated worse

Indie (smaller) rated better

(-2) QTY GENRES

Specialization is preferred

Meshing genres may lead to poorer game design

Misleading consumers

(0) GAME PRICE

Rating is minimally effected by pricing

Little incentive to sell at lower prices

Microtransactions profitable

(N) FEATURE: where N is **β**-coefficient rounded to nearest integer

FUTURE WORK

1

BALANCE DISTRIBUTION

Data is heavily skewed;
only 12% of games rated
between 0-50%

2

REGRESSION MODELS

Other models
(*e.g.* polynomial regression or
different distributions)
for better modeling

3

SUCCESS METRIC

Look at active players or
recurring sales

4

ADDITIONAL FEATURES

Wider scrape of features
that influence video
game success

Thank You

Does anyone have any questions?

Brian Nguyen
at Metis

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**.

Please keep this slide for attribution.

APPENDIX

TABLE OF CONTENTS

01 **BACKGROUND**
Target: Overall User Rating

02 **DATA OVERVIEW**
Scraping, Cleaning, and Feature Engineering

03 **FEATURES**
Significant and Insignificant Features

04 **MODELING**
Finding the Most Appropriate Model

05 **REGULARIZATION**
Examining Ridge versus LASSO

06 **FUTURE WORK**
Notable Concerns and More Plans (Given Time)

GENERATING DATA FRAME

WEB SCRAPING

store.steampowered.com

Generate links for all games

Scrape game details
from specific product link

DATA CLEANING

Set an **'active game' criterion**:
minimum 1 review/day (average)

Remove unreleased games

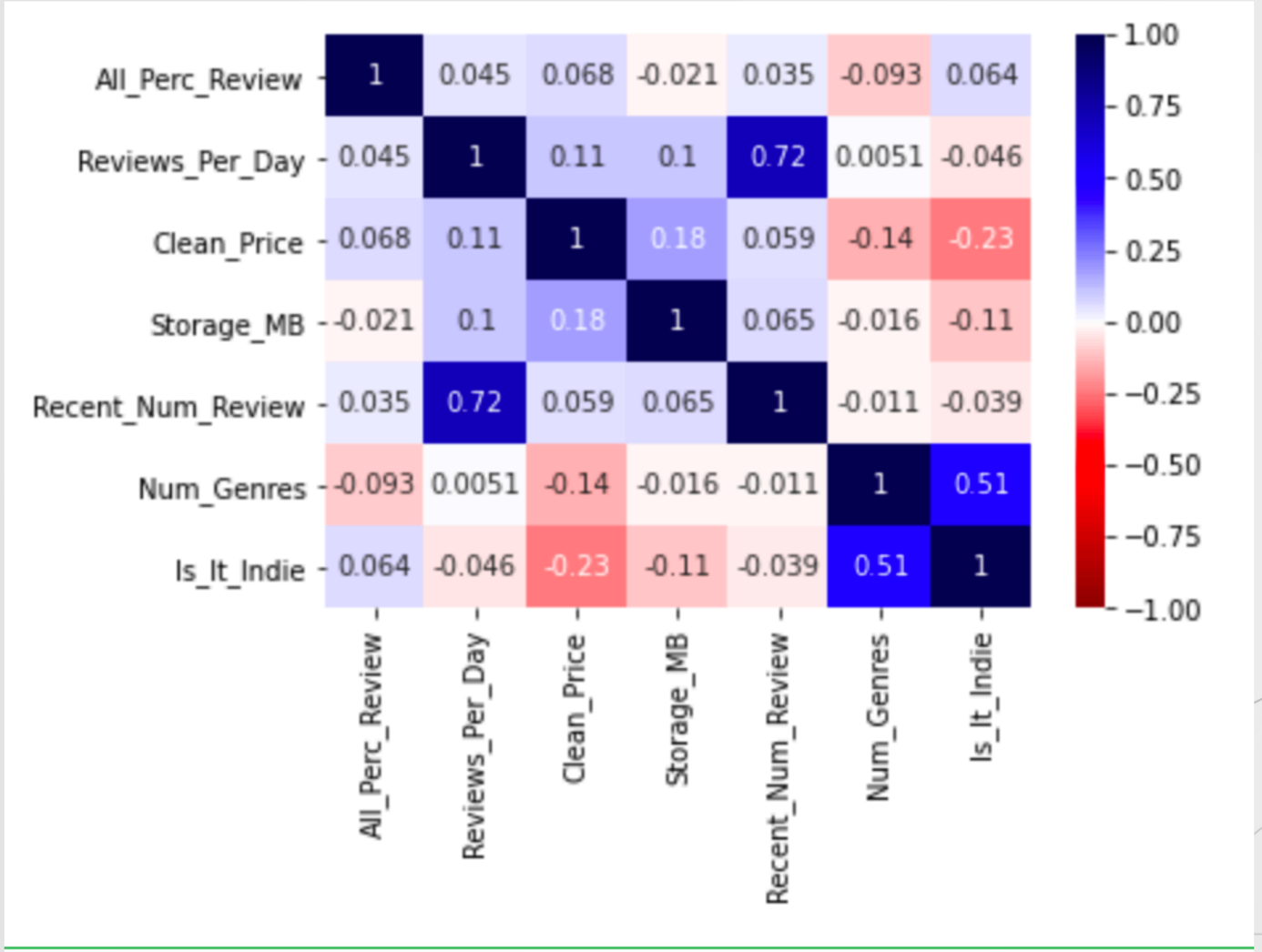
Remove bundles (> $60)

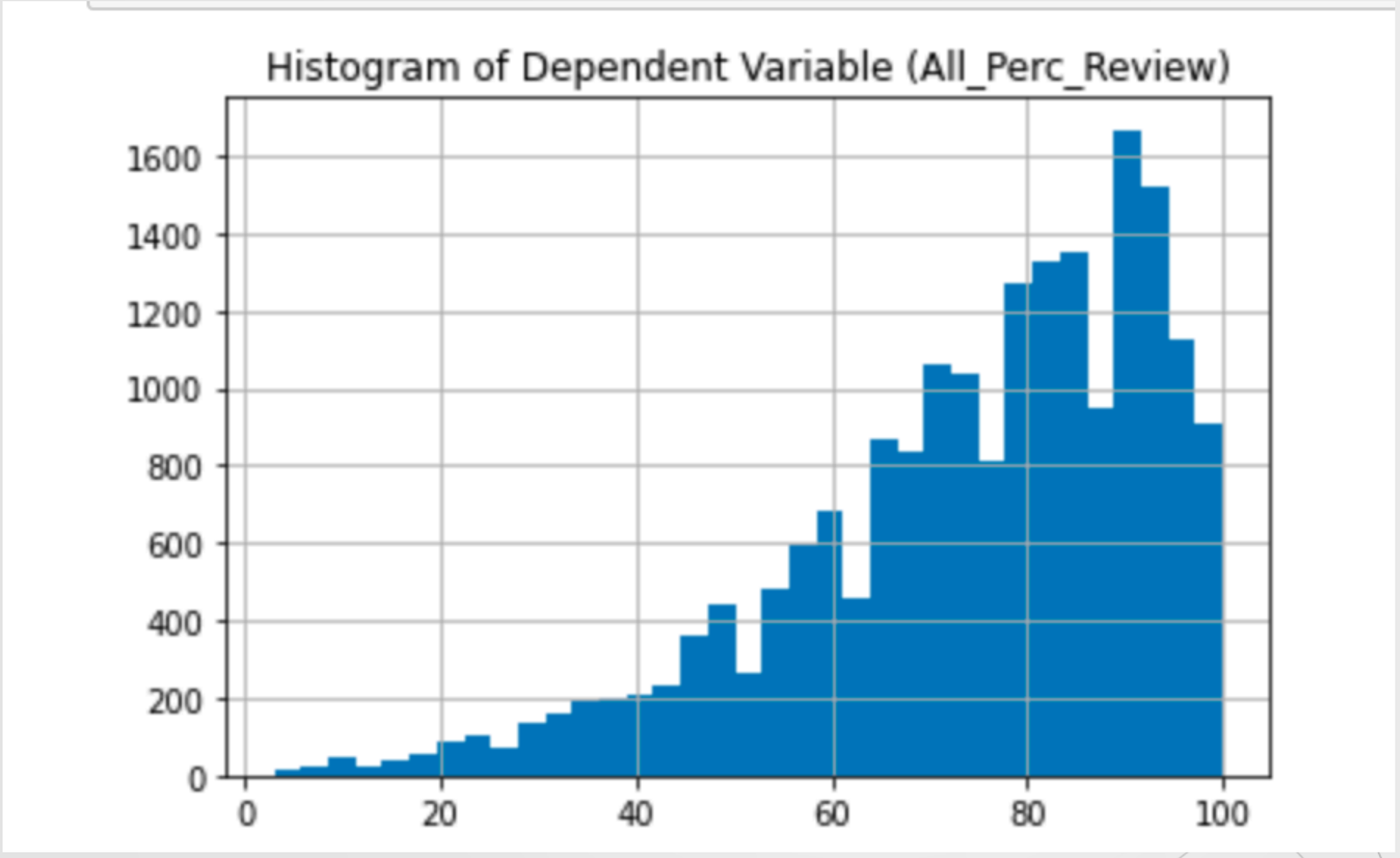
FEATURE ENGINEERING

Normalize all number reviews into
reviews/day

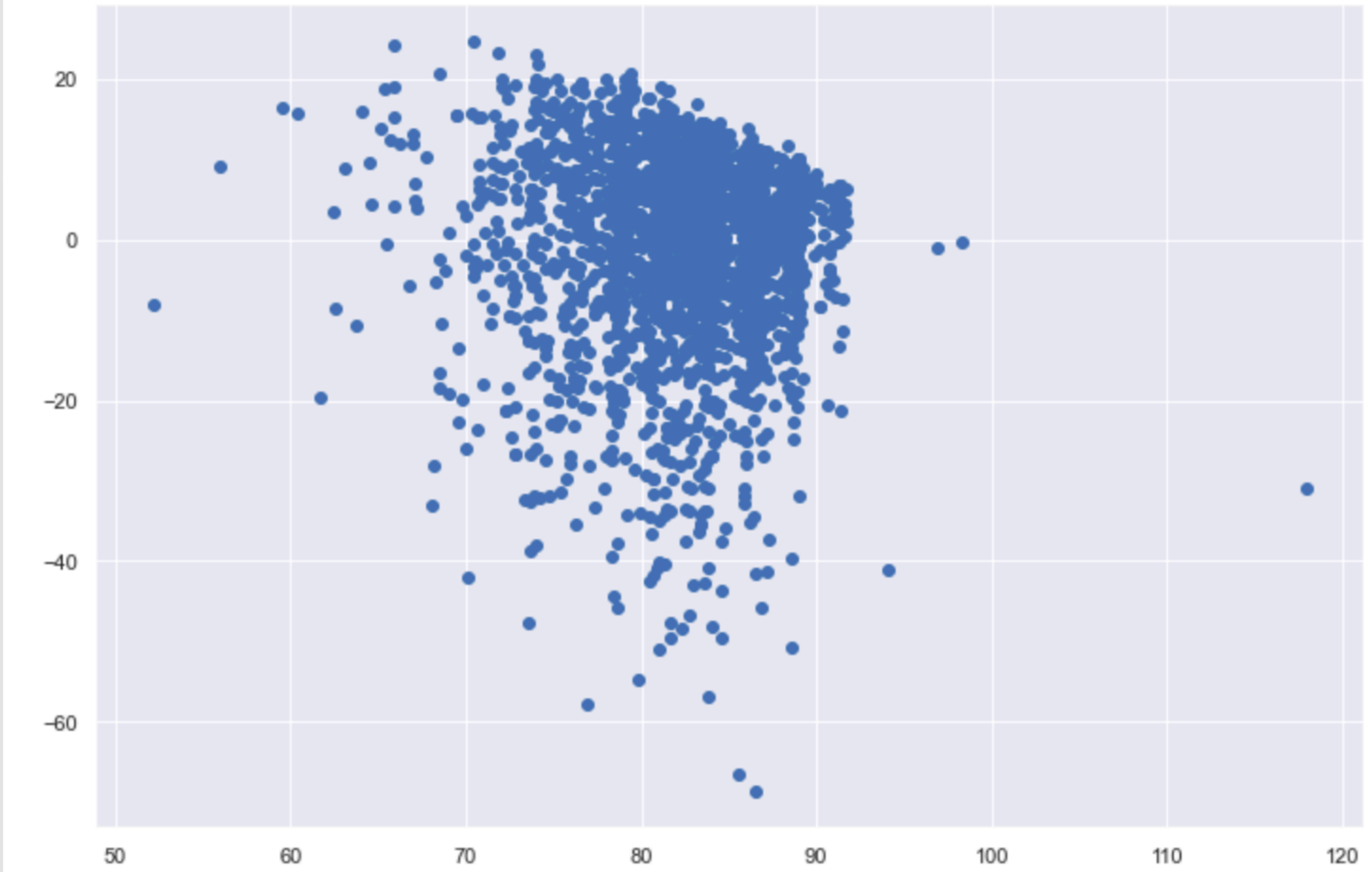
Convert list of genres to quantity

Categorize independent
developer (indie) as yes / no


All_Perc_Review
Reviews_Per_Day
Clean_Price
Storage_MB
Recent_Num_Review
Num_Genres
Is_It_Indie
1 0.045 0.068 -0.021 0.035 -0.093 0.064
0.045 1 0.11 0.1 0.72 0.0051 -0.046
0.068 0.11 1 0.18 0.059 -0.14 -0.23
-0.021 0.1 0.18 1 0.065 -0.016 -0.11
0.035 0.72 0.059 0.065 1 -0.011 -0.039
-0.093 0.0051 -0.14 -0.016 -0.011 1 0.51
0.064 -0.046 -0.23 -0.11 -0.039 0.51 1
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

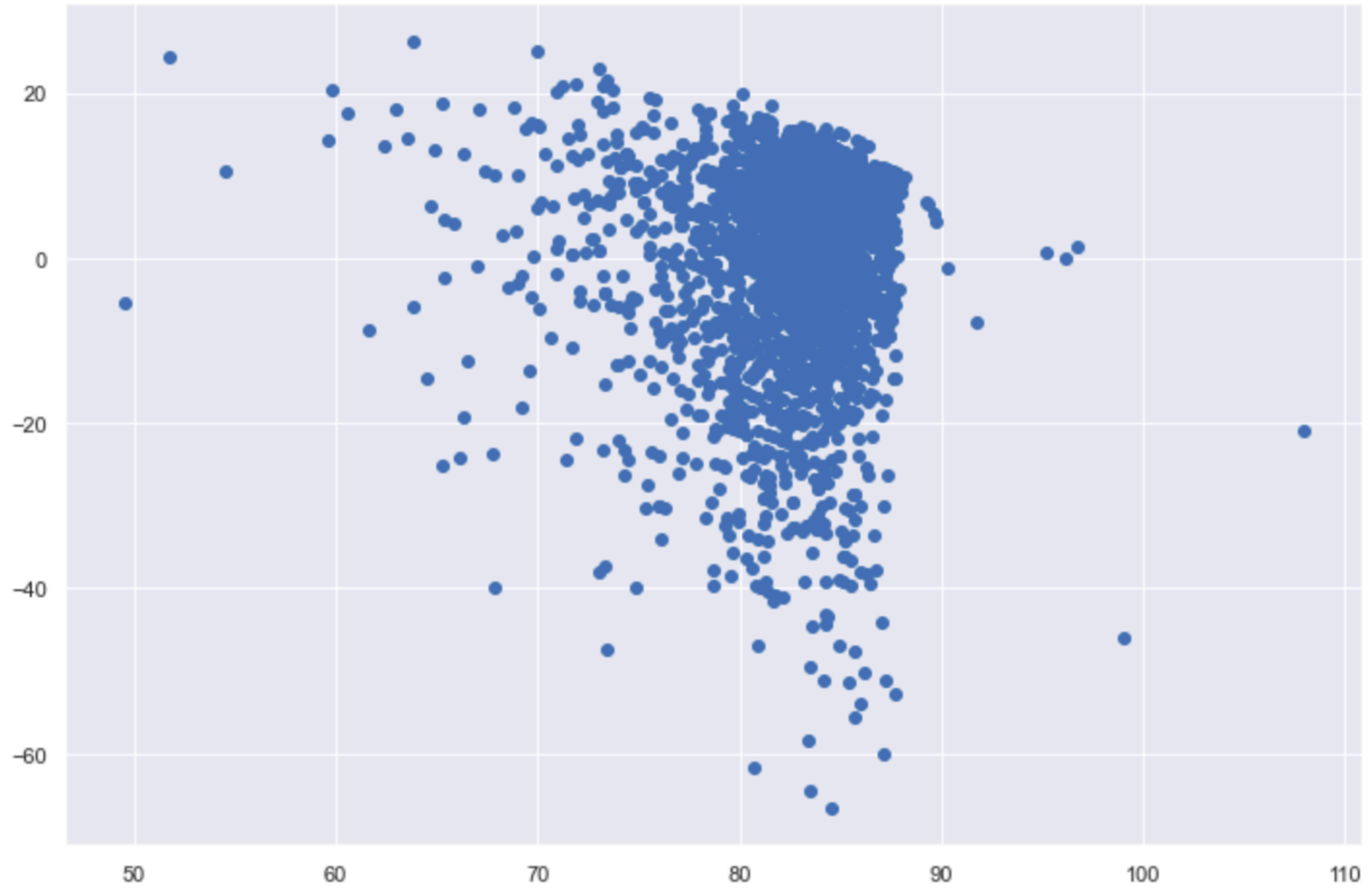

Histogram of Dependent Variable (All_Perc_Review)
1600
1400
1200
1000
800
600
400
200
0
0
20
40
60
80
100

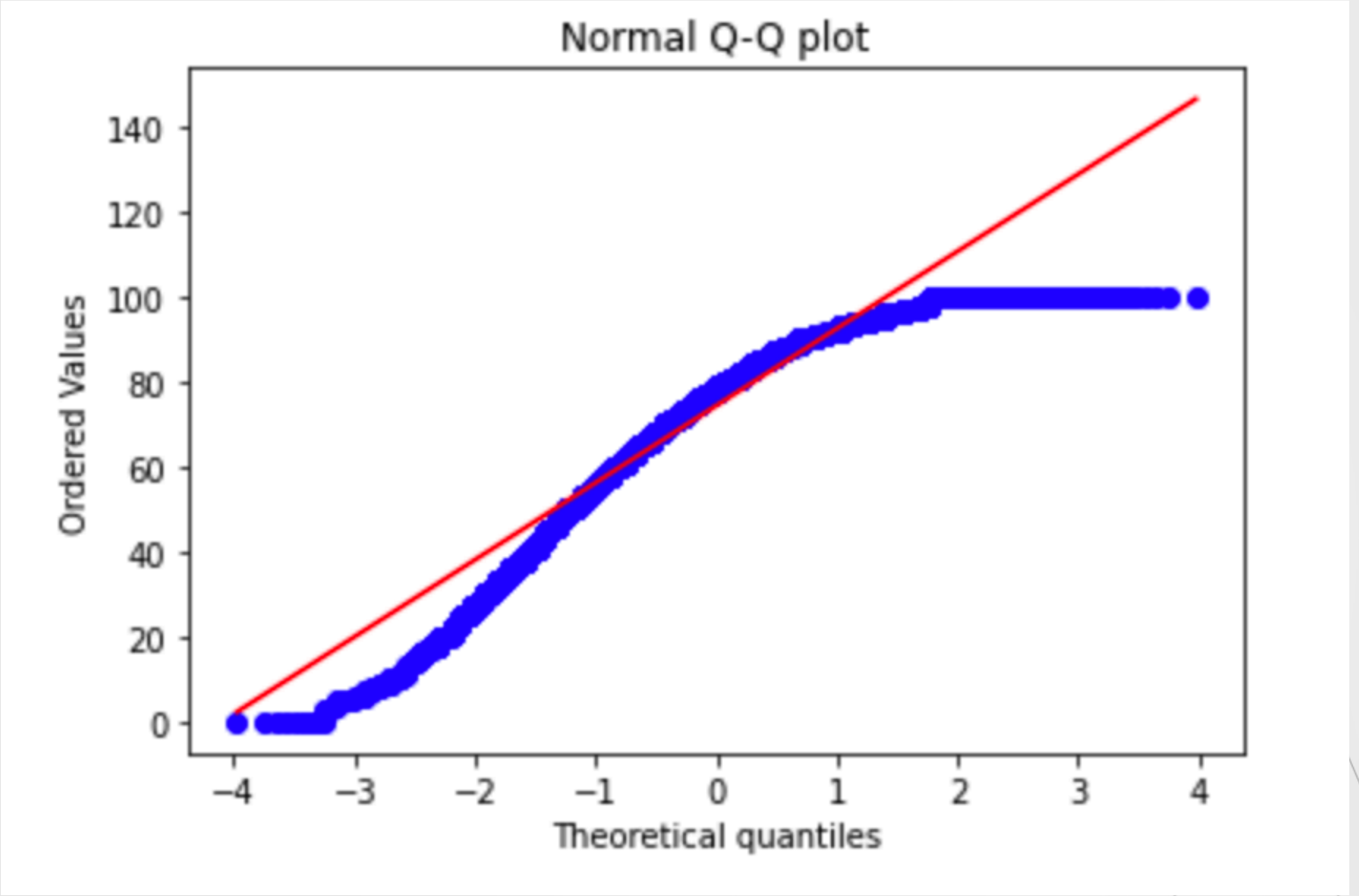
```
1  # Residual plot with 6 features
2  plt.figure(figsize=(12,8))
3  plt.scatter(fit_overview.predict(), fit_overview.resid);
```



In [20]:

```
1  # Residual plot with 4 features
2  plt.figure(figsize=(12,8))
3  plt.scatter(fit_overview.predict(), fit_overview.resid);
```




Normal Q-Q plot
Ordered Values
Theoretical quantiles
140
120
100
80
60
40
20
0
−4
−3
−2
−1
0
1
2
3
4