

---

# ReCell

Eugenie Seholm

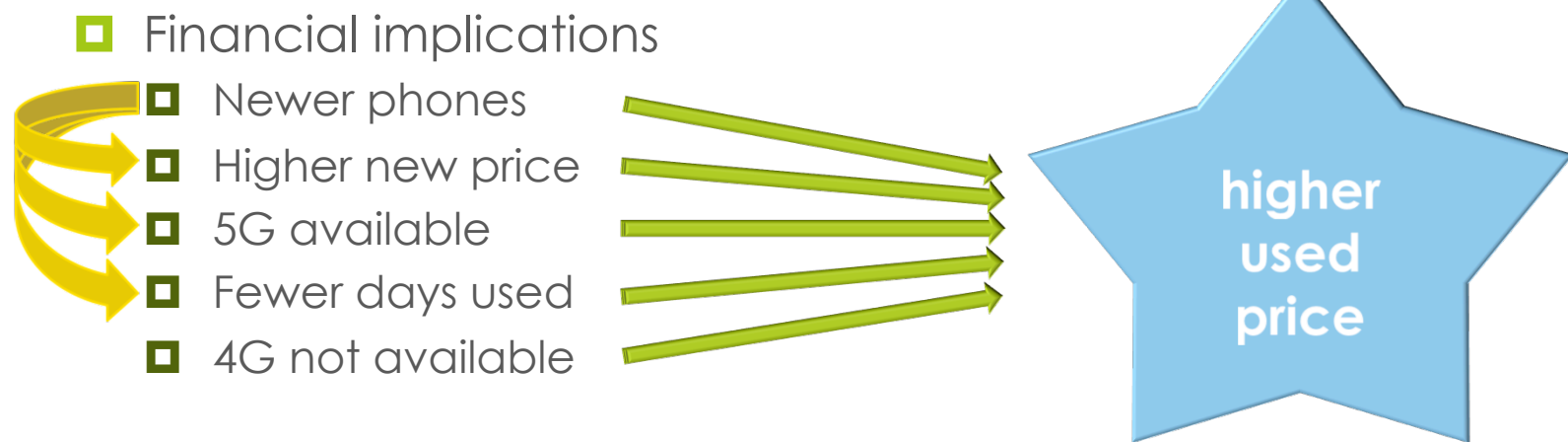
---

# Contents

- Business Problem Overview and Solution Approach
  - Data Overview
  - Data Manipulations
  - Exploratory Data Analysis (EDA)
  - Model Performance Summary
  - Business Insights and Recommendations
-

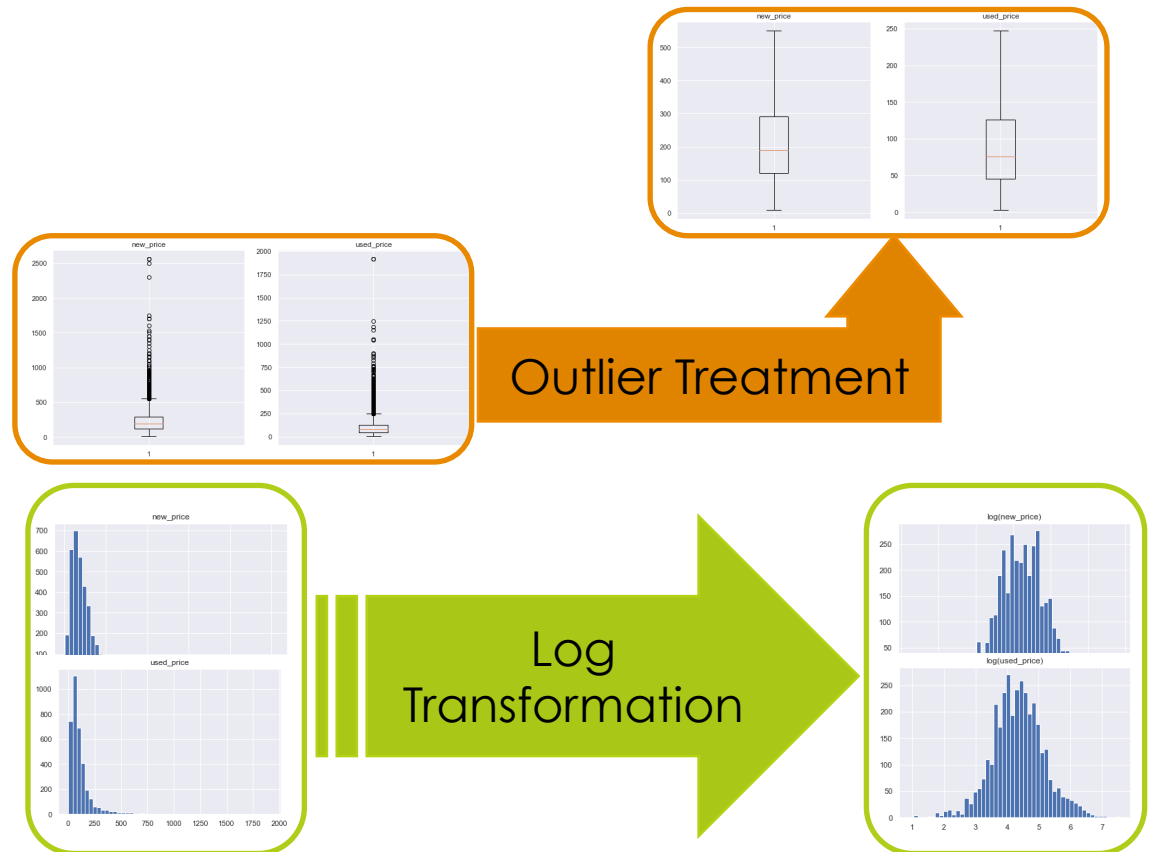
# Business Problem Overview and Solution Approach

- Core business idea:
  - Create a pricing strategy for used and refurbished phones
- Problem to tackle:
  - Identify key factors that significantly influence used price and build a linear regression model



# Business Problem Overview and Solution Approach

- ML model to solve the problem
  - Fit linear model to existing data
  - **Treat outliers** and fit linear model to data
  - **Log transform** original data and fit linear model
  - Linear model on log transformed data has best performance



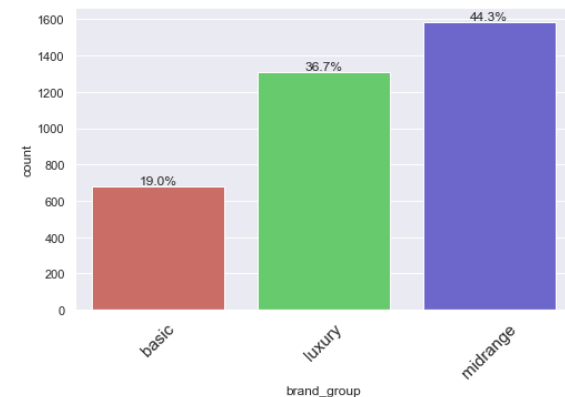
# Data Overview

- 3571 observations (rows)
- 15 variables (columns)
  - 4 categorical
  - 11 numerical
- Missing values (to impute):
  - main\_camera\_mp: 180
  - selfie\_camera\_mp: 2
  - int\_memory: 10
  - ram: 10
  - battery: 6
  - weight: 7

Variable	Description
brand_name	Name of manufacturing brand
os	Operating system
screen_size	Size of screen (cm)
4g	Whether 4g is available or not
5g	Whether 5g is available or not
main_camera_mp	Resolution of rear camera (megapixels)
selfie_camera_mp	Resolution of front camera (megapixels)
int_memory	Amount of internal memory (ROM) in GB
ram	Amount of RAM in GB
battery	Energy capacity of phone battery (mAh)
weight	Weight of phone (grams)
release_year	Year when phone model was released
days_used	Number of days the used/refurbished phone was used
new_price	Price of a new phone of the same model (euros)
used_price	Price of the used/refurbished phone (euros)

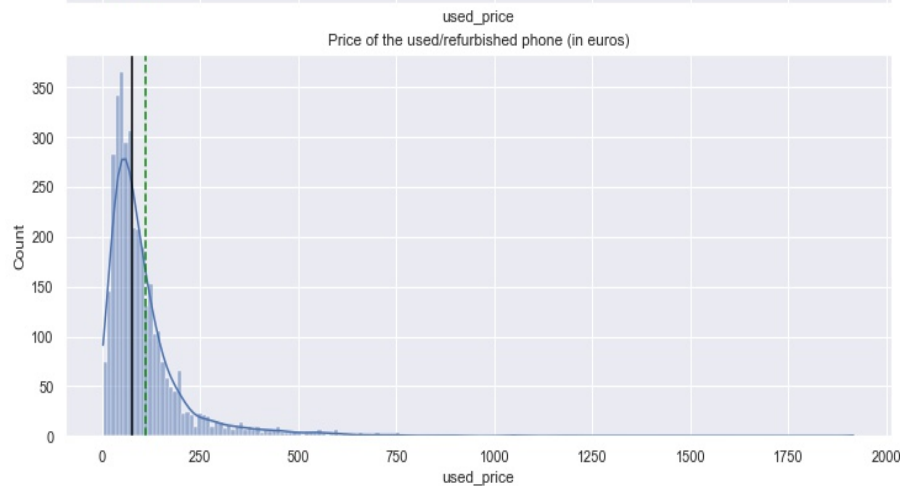
# Data Manipulations

- Brand name grouped by brand's mean price
  - "Basic" – cheapest 12 brands
  - "Midrange" – middle 11 brands
  - "Luxury" – priciest 11 brands
- Missing values
  - Predictor variables all numerical
  - Replace with median for each column
- Bin *main\_camera\_mp* and *selfie\_camera\_mp* for visualization
  - poor: up to 2 MP
  - good: 2 to 8 MP
  - great: 8 to 12 MP
  - excellent: more than 12 MP
- Outlier treatment on *new\_price* and *used\_price*
- Log transformation on *new\_price* and *used\_price*



# EDA: Used Price (Target)

- Positively skewed data
- Outliers in positive direction
- Some extreme outliers

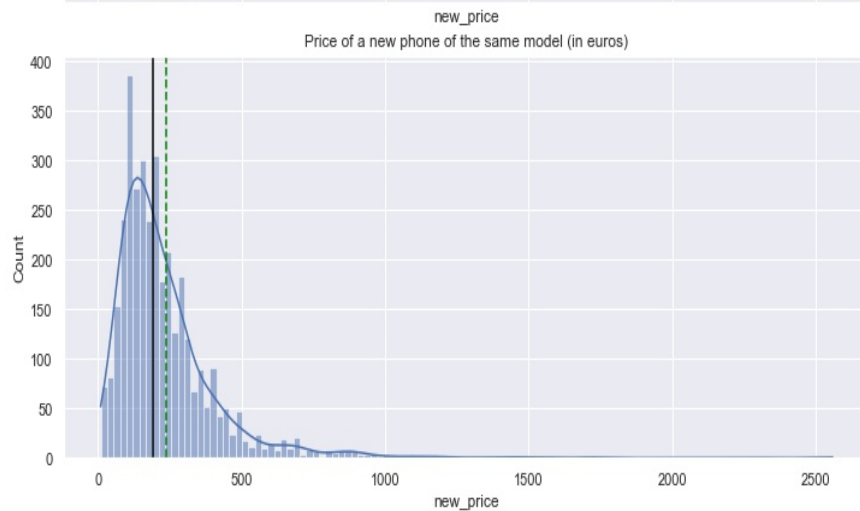


- Median: 75.5 euros
- Range: 2.5 – 1916.5 euros
- Log transformation much closer to normal distribution



# EDA: New Price (significant factor)

- Positively skewed data
- Outliers in positive direction
- Some extreme outliers



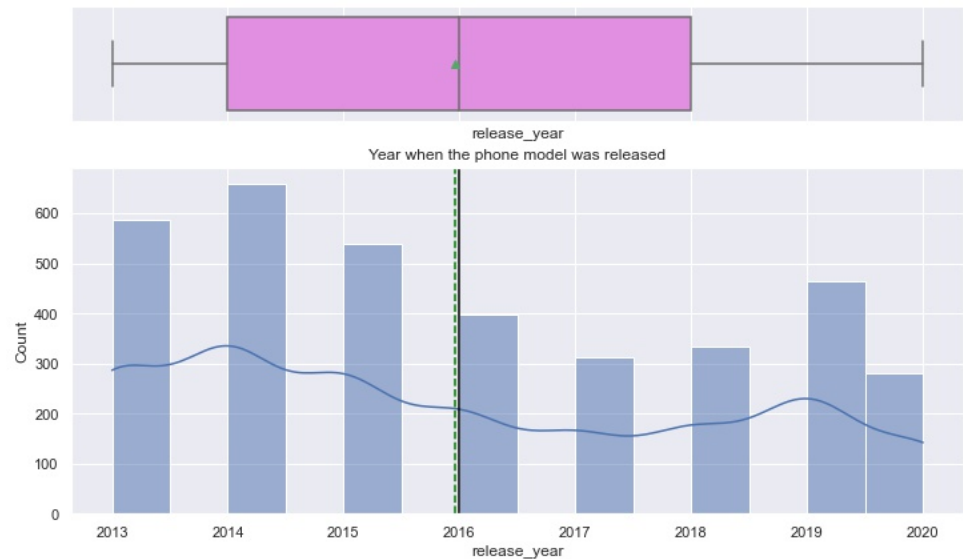
- Median: 189.8 euros
- Range: 9.13 – 2560.2 euros
- Log transformation much closer to normal distribution



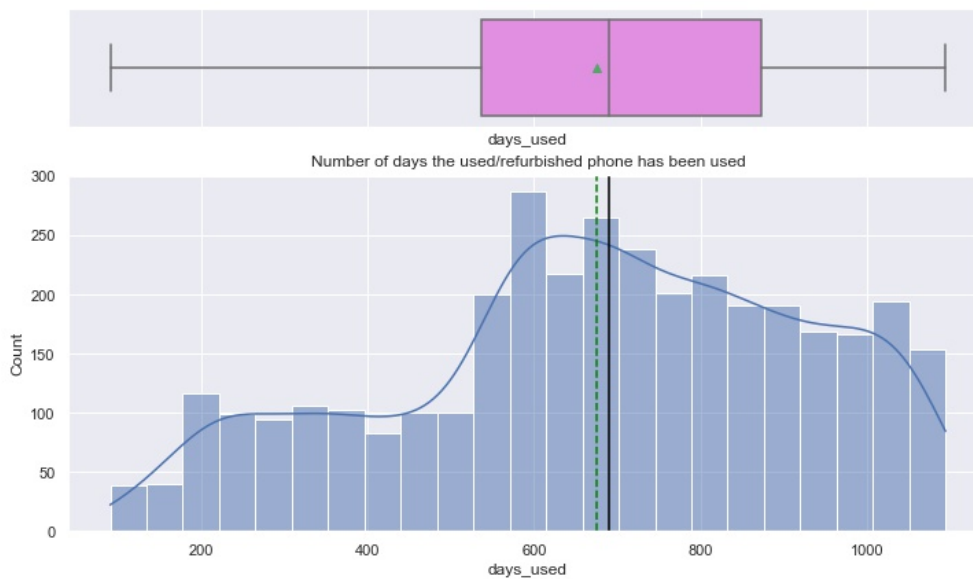


# EDA: Release Year (significant factor)

- Bimodal appearance
- Slight negative skew
- No outliers
- 2013 – 2020
- Mode: 2014
- Median: 2016



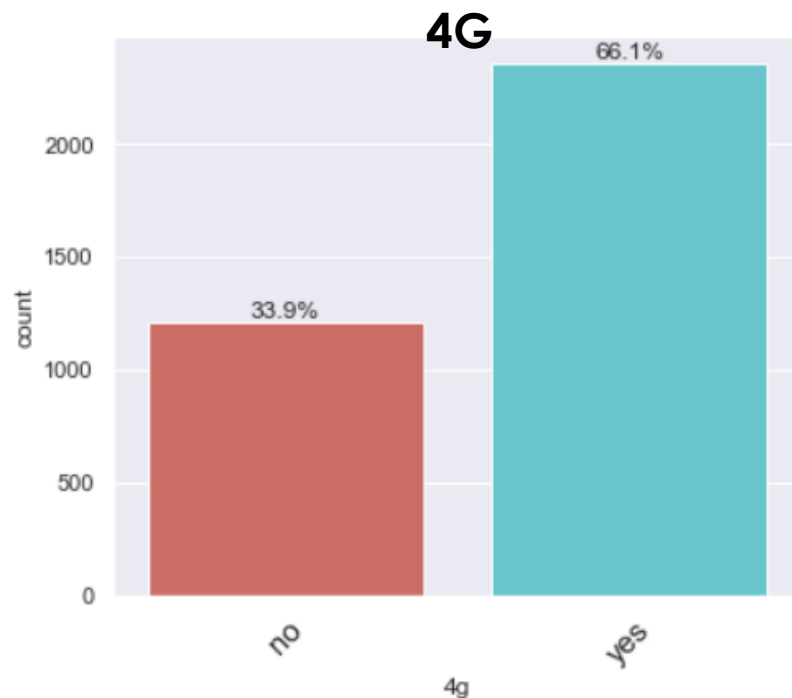
# EDA: Days Used (significant factor)



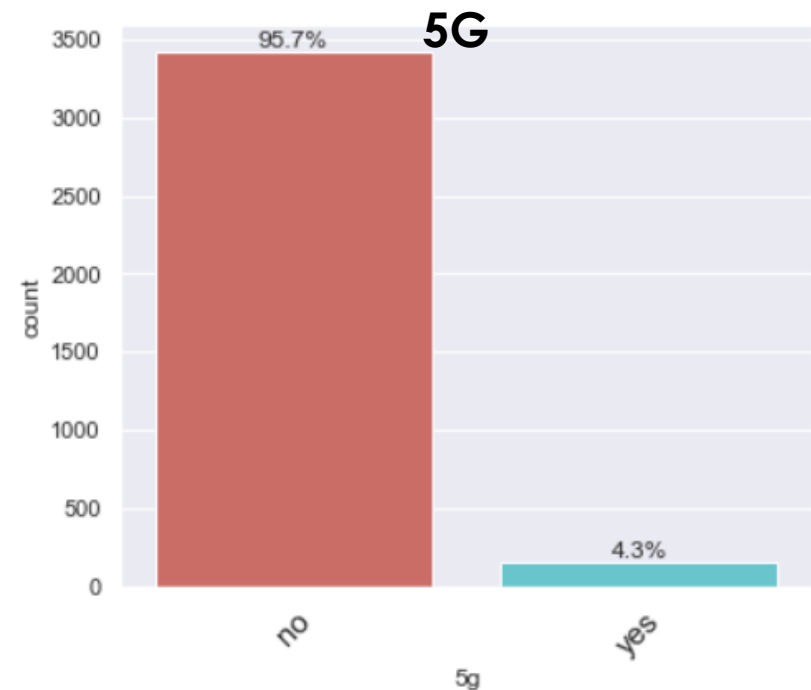
- Slight negative skew
- No outliers
- Median: 690 days
- Mean: 675 days
- Range: 91 – 1094 days (3 months – 3 years)

# EDA: 4G & 5G (significant factors)

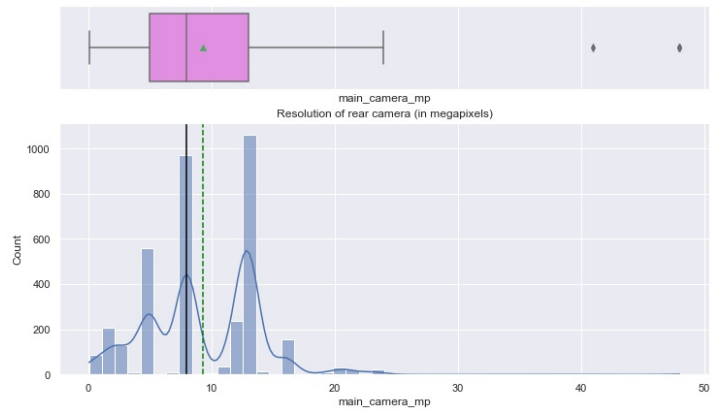
■ Majority (66.1%) have 4G



■ Majority (95.7%) do NOT have 5G

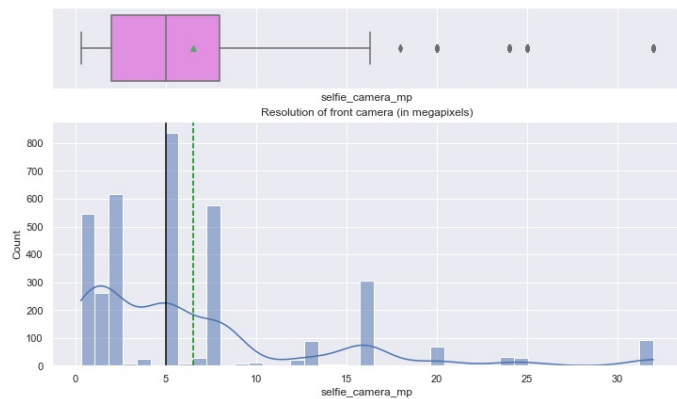
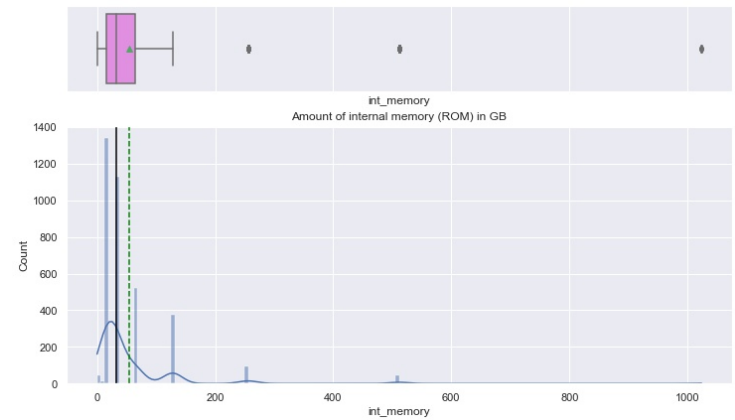


# EDA: Main & Selfie Cameras, Memory

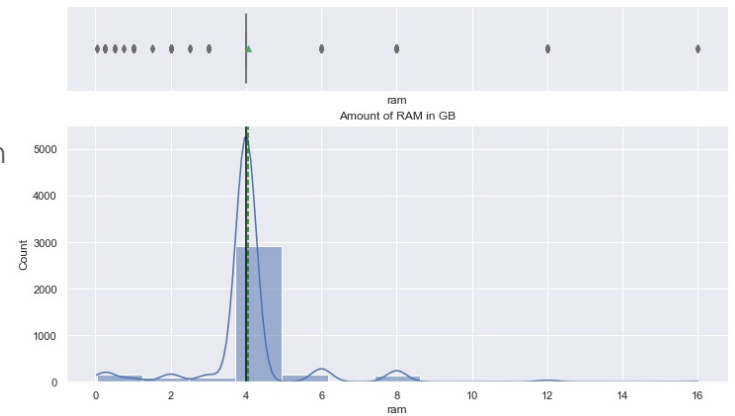


- Main & Selfie Cameras
  - Positive skew
  - Outliers in positive direction

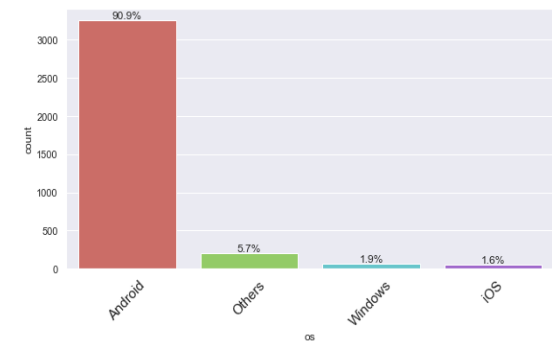
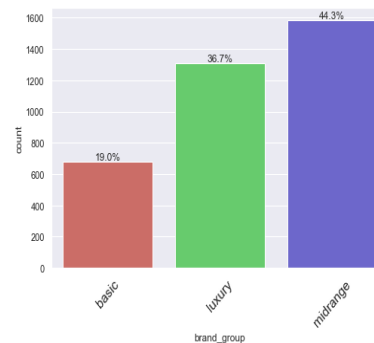
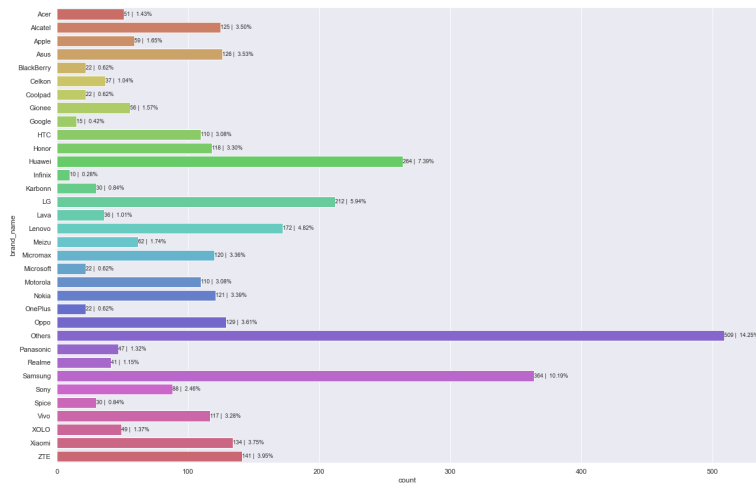
- Internal Memory
  - Positive skew
  - Outliers in positive direction
  - Some extreme outliers



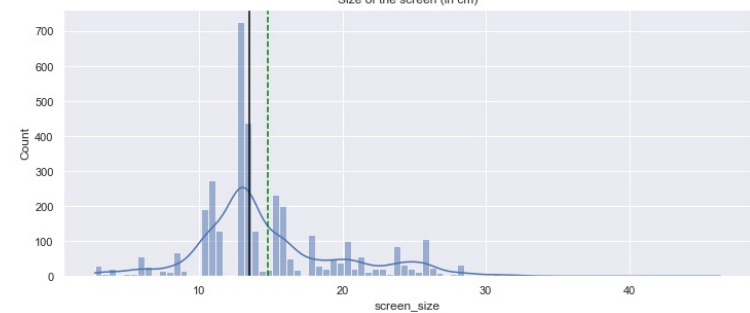
- RAM
  - Positive skew
  - Outliers in both directions
  - Most of the data has 4 GB RAM
  - Some extreme outliers



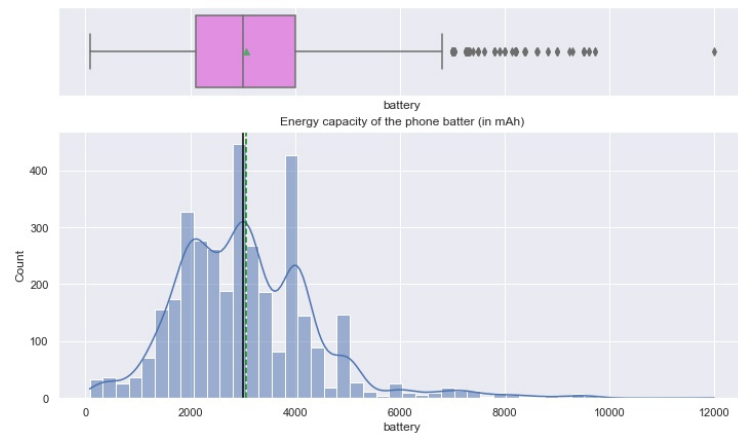
# EDA: Brand Name, OS, Screen Size



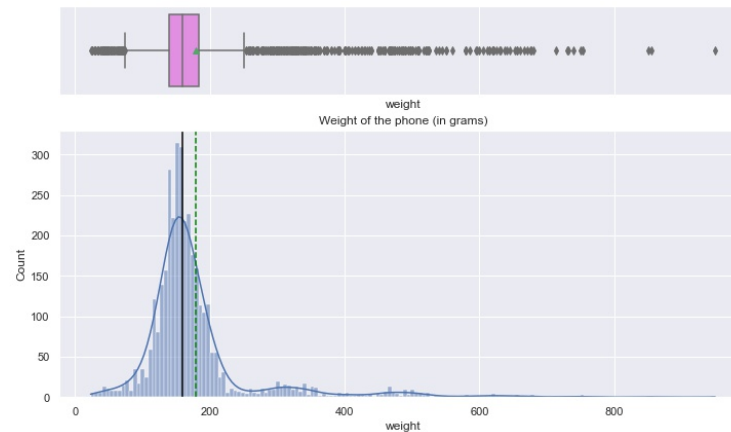
- Samsung and Huawei are most popular brand names
- Most phones are midrange
- Android os dominates used phone market
- Screen size positively skewed with outliers in both directions, and some extreme outliers
- Some screen sizes do not seem plausible



# EDA: Battery, Weight



- Battery
  - Positively skewed
  - Outliers in the positive direction
  - Appears multi-modal

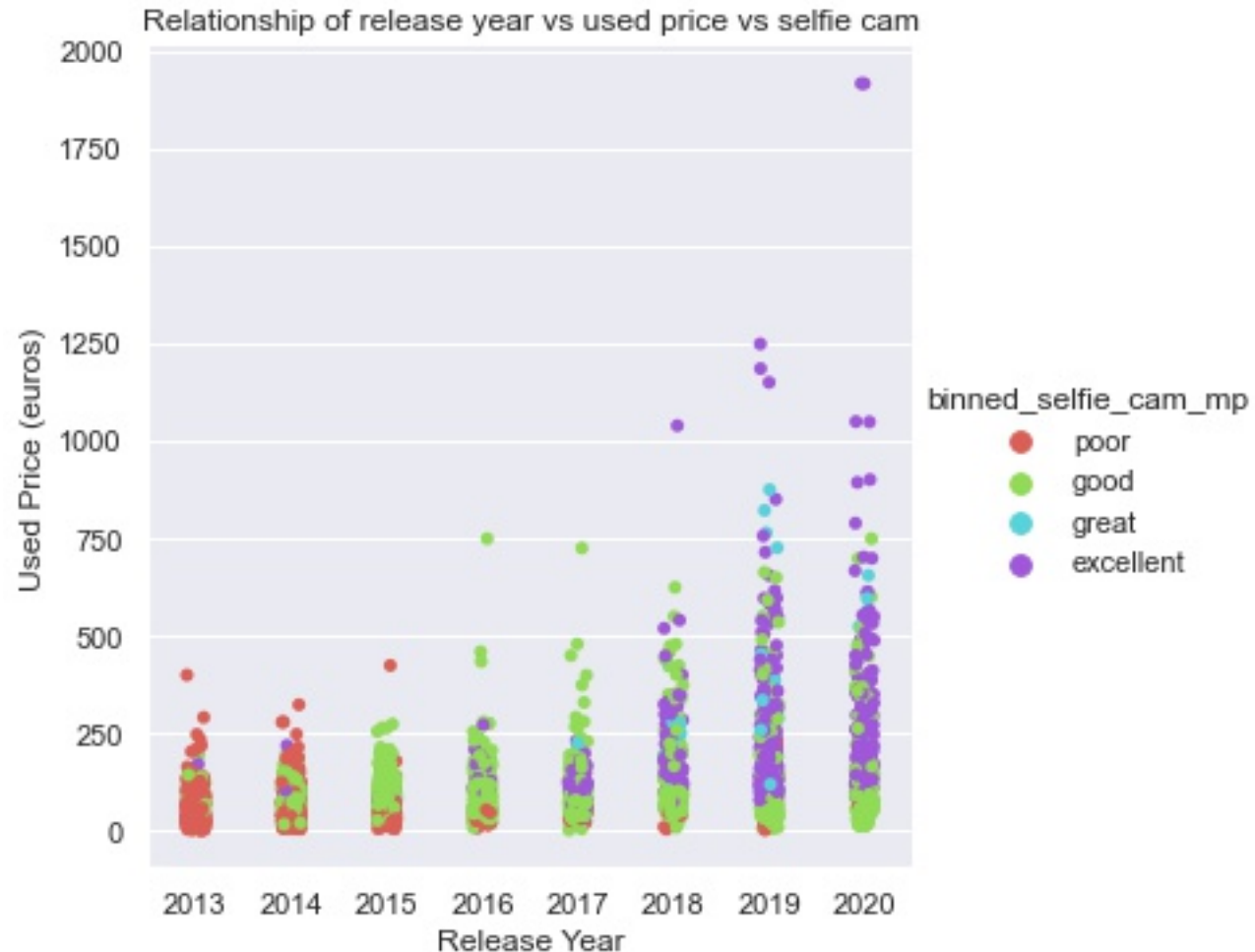


- Weight
  - Positively skewed
  - Outliers in both directions
  - Extreme outliers in positive direction (by error?)



## EDA: Year vs New/Used Price

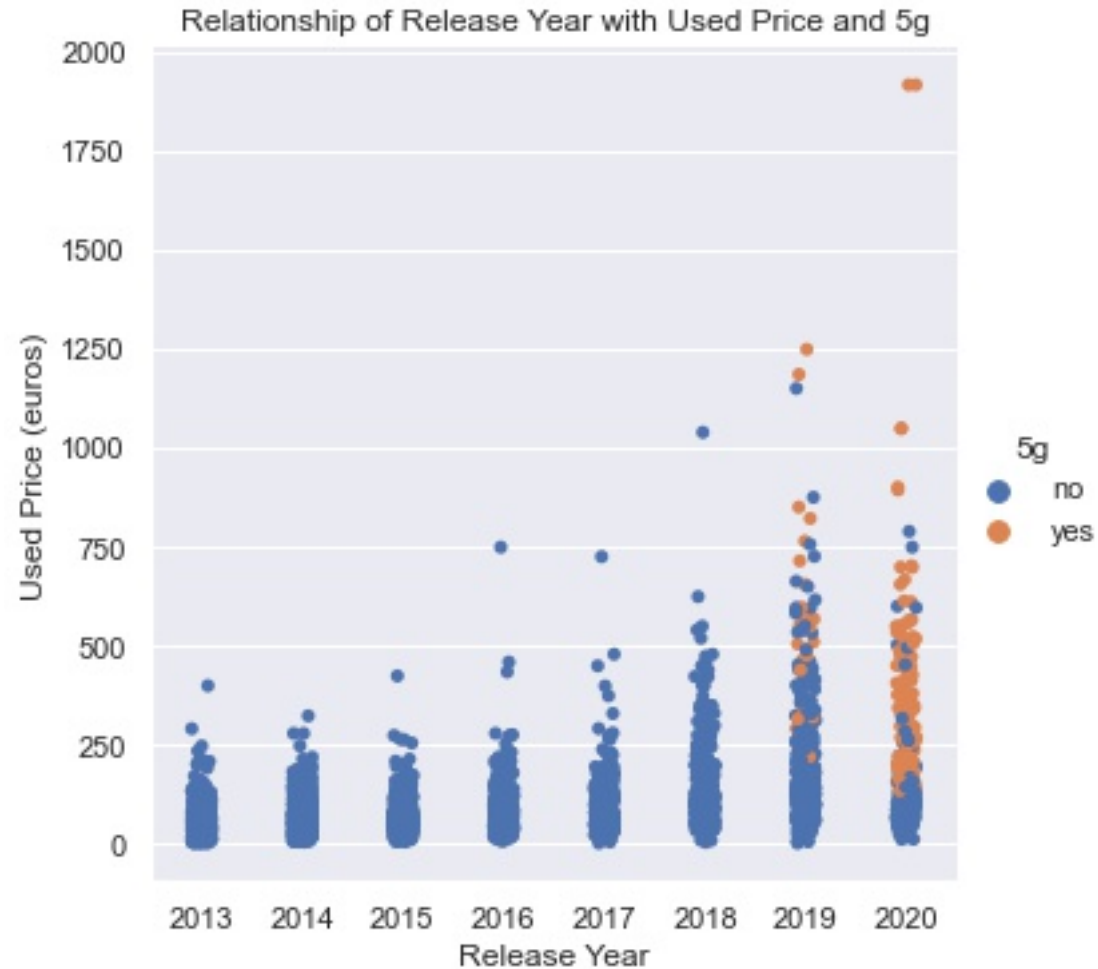
For any given new price, newer phones yield higher used price



## EDA: Year / Used Price / Selfie Cam

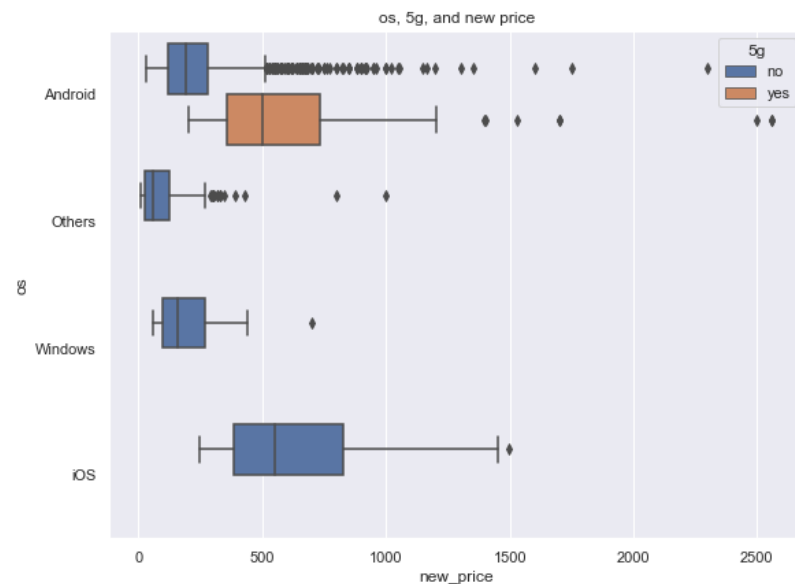
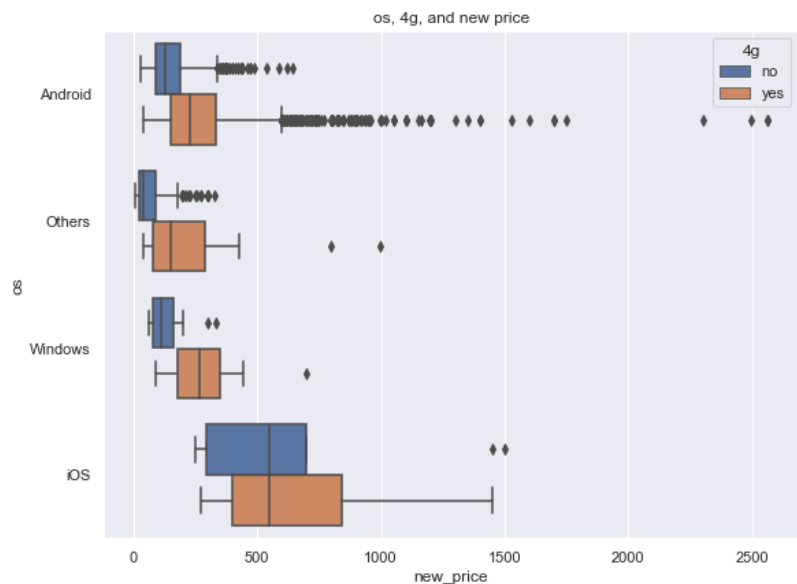
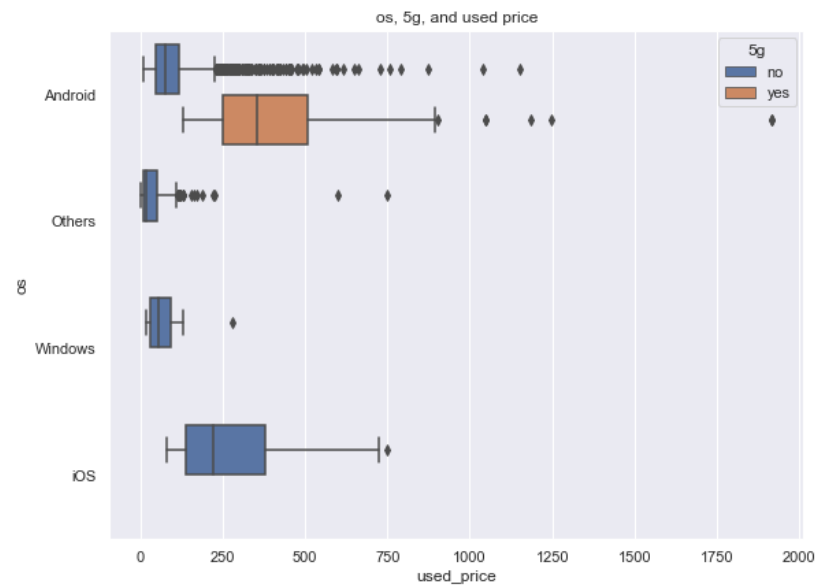
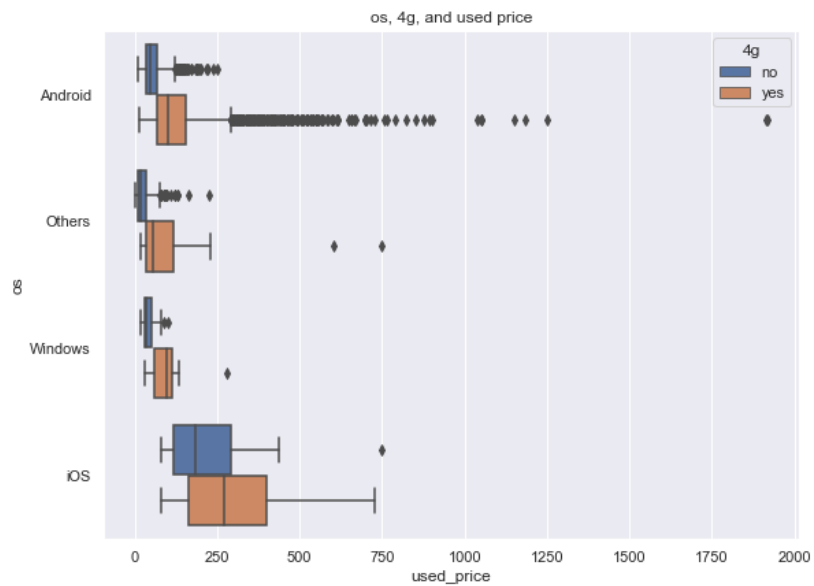
Newer phones tend to have better selfie cameras, which yield higher used price





## EDA: Year / 5G / Used Price

Only newer phones can have 5G and 5G yields a higher used price



- highest new & used prices: iOS (higher prices for 4G; no 5G)
- 2<sup>nd</sup> highest new & used prices: Android (higher prices for 4G and for 5G)
- 3<sup>rd</sup> highest new & used prices: Windows (higher prices for 4G; no 5G)
- lowest new & used prices: Others (Higher prices for 4G; no 5G)

os **vs** 4G or 5G **vs**  
new or used price

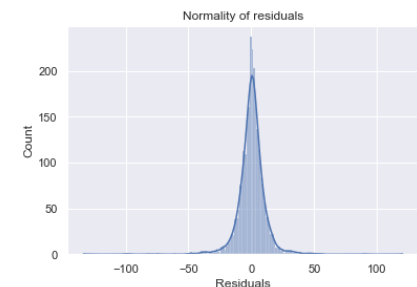
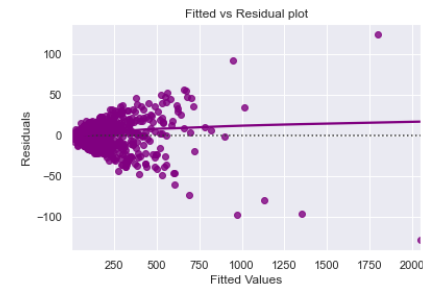
# Model Performance Summary

- statsmodels least squares
- log-transformed variables:
  - target variable  
*used\_price*
  - independent variable  
*new\_price*
- brands grouped by brand name's mean price
- categorical variables with one-hot encoding
- variables with  $p > 0.05$  dropped one by one in order of highest  $p$ -value and then created new model, checking  $p$ -values each time
- no multicollinearity
- assumption of linearity & independence satisfied
- assumption of normality of error terms satisfied
- assumption of homoscedasticity of residuals satisfied

# Model Performance Summary

- variance inflation factors (VIF) for all variables between 1 and 5 so we have **low multicollinearity**
- assumption of **linearity and independence satisfied** because we see no pattern in fitted vs residuals plot
- assumption of **normality of error terms satisfied** since residuals have bell shape and we can accept the distribution as close to normal as an approximation
- The model works fine for most of the data and Goldfeld-Quant test indicates heteroskedasticity, but it may be due to limited data so we accept the distribution of residuals to satisfy the condition of **homoscedasticity**

	feature	VIF
0	const	2.513334e+06
1	release_year	3.257568e+00
2	days_used	2.446705e+00
3	new_price_log	1.401414e+00
4	4g_yes	1.925431e+00
5	5g_yes	1.260972e+00

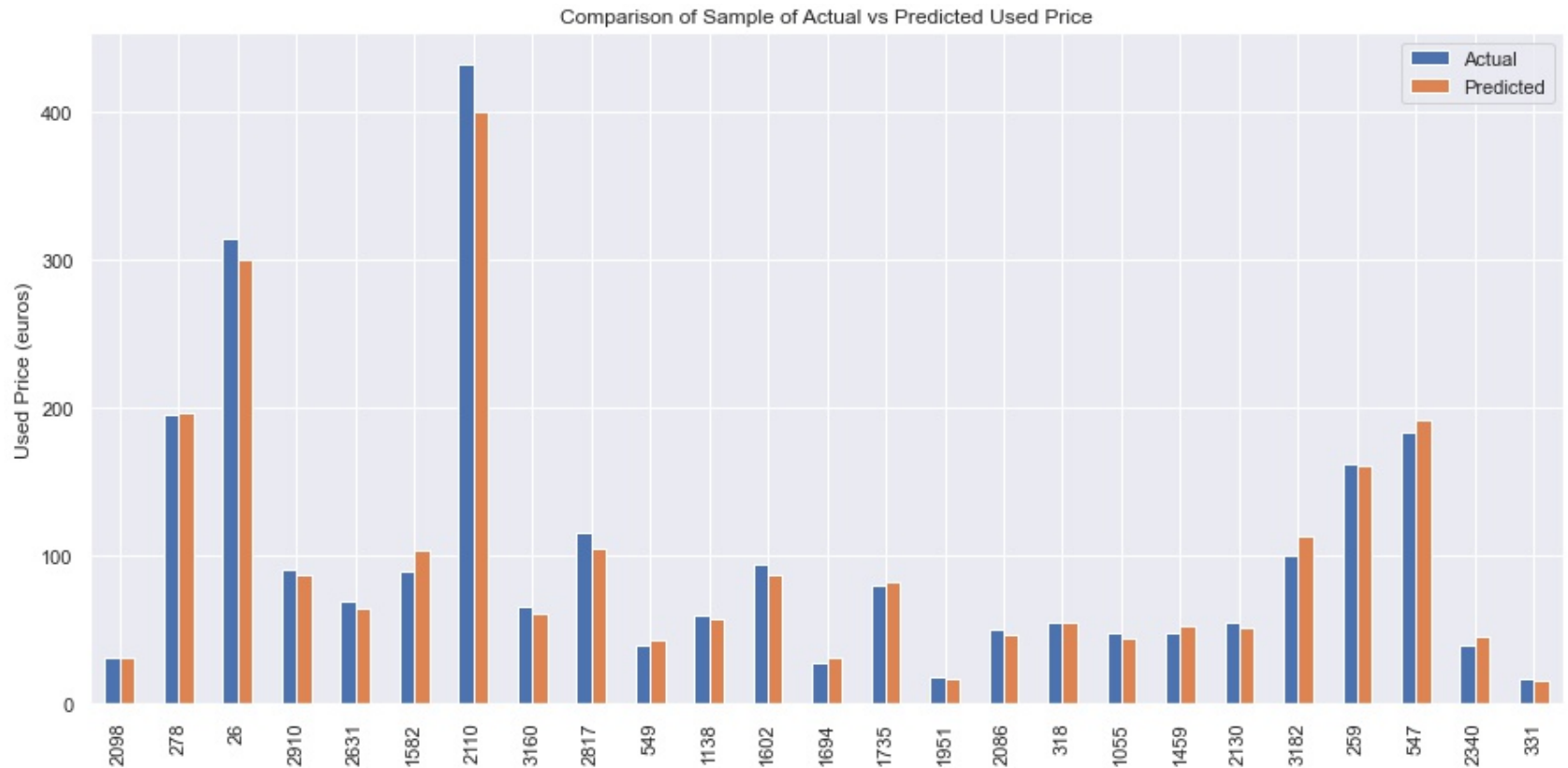


# Model Performance Summary

Variable	Coefficient
constant	-12.925069
release_year	0.006349
days_used	-0.001111
new_price_log	1.000773
4g_yes	-0.010437
5g_yes	0.024723

## SIGNIFICANT VARIABLES

- As **new price** increases by 1 euro, used price **increases by ~0.9%**
- As **release year** increases by 1 year, used price **increases ~0.63%**
- As **days used** increases by 1 day, used price **decreases ~0.11%**
- Phones with **4g** have used price **~1.04% less** than those without
- Phones with **5g** have used price **~2.47% more** than those without



# Model Performance Summary

Our model has returned good prediction results and the actual values are comparable to predicted values

# Model Performance Summary

	Training Performance	Test Performance
<b>RMSE</b>	11.329552	11.557044
<b>MAE</b>	7.022132	7.291537
<b>R<sup>2</sup></b>	0.991625	0.990057
<b>Adjusted R<sup>2</sup></b>	0.991605	0.990001
<b>MAPE</b>	7.005776	7.07044

- Model is able to explain ~99% variation in data
- Train and test RMSE and MAE are low and comparable
- MAPE: We can predict within 7.1% of the used price

# Business Insights and Recommendations

- Factors that influence used price the most are release year, number of days used, new price, and availability of 4G and/or 5G.
  - Newer phones tend to have higher used price for any given new price
  - Only newer phones have 5G available (although many do not have it yet), which results in higher new and used prices
  - For any given release year, the more a phone is used, the lower its used price



# Business Insights and Recommendations

- Based on EDA, popular phones have:
  - Brand Samsung & Huawei
  - Android operating system
  - screen size 5" – 6.5"
  - 4G signal strength
  - 5 – 13 megapixels main camera
  - 2 – 8 megapixels selfie camera
  - 4 GB ram
  - weight 140 – 184 grams
  - at least 3 years old
  - new price under €292
  - used price under €126

# Business Insights and Recommendations

- For pricing used/refurbished phones focus on
  - new price
  - signal strength (4G/5G)
  - release year
  - days used
- Use pricing information to determine
  - reasonable buyback price
  - competitive sale price
    - basic phones: most affordable – likely smallest margin
    - midrange phones: most available
    - luxury phones: highest used price – likely largest margin

# Business Insights and Recommendations

- Note that different cameras (both main & selfie) take vastly different quality photos, even if they have the same number of megapixels
  - Different sensors
  - Different number/type of lenses
  - Placement of lenses
- Collect more data, as it may impact our model. As technology improves, we will likely see some trends, so with more data, we will likely start to see more information on categories we currently lack:
  - Newer phones
  - Greater number of days used
  - 4G signal strength availability
  - 5G signal strength availability
  - Luxury brand phones
  - Lightweight phones
  - Less expensive brand phones
  - Higher battery capacity
  - Better camera quality

# Business Insights and Recommendations

Collect more data on different variables. There may be additional factors that influence used phone price:

- ▣ Color(s) available
- ▣ Material
- ▣ Thickness
- ▣ Networks
- ▣ SIM (single/dual, etc.)
- ▣ Display type
- ▣ Resolution
- ▣ Memory card slot
- ▣ Video frames per second
- ▣ Camera lenses/sensors
- ▣ Speakers
- ▣ 3.5 mm jack
- ▣ Bluetooth
- ▣ USB
- ▣ Battery type
- ▣ Fast charging available
- ▣ Hotspot capability
- ▣ GPS
- ▣ Radio