An Overview of the Irish Car Market with Cars Priced Under 7000 Euros Between 2023 and 2024

Attila Benczik

Table of Contents

Abstract

This report investigates price variations in cars priced under €7000 between 2023 and 2024 and develops a model to accurately predict car prices based on key predictors, such as mileage, age, model, and year. A Mixed Linear Model (MLM) was employed to achieve this, demonstrating the ability to predict prices for most car models with high accuracy. The mixed-effects model was selected for its capacity to account for both the fixed effects of general trends across the dataset and the random effects associated with individual car models. The findings suggest the model provides reliable price predictions, making it a valuable tool for prospective car buyers, sellers, and dealerships.

1 Chapter 1: Introduction

1.1 Background and motivation

This report analyses the Irish car market, focusing specifically on vehicles priced under €7,000. By examining data within this affordable segment, the analysis aims to understand and predict car prices in this price range. Key factors considered include car models and production years, with the objective of helping Irish car buyers identify good-value cars. This analysis is intended to support informed decision-making and potentially save buyers money in challenging economic times.

The report also aims to provide valuable information for "car flippers", individuals who buy used cars with the intent to resell at a profit. By identifying vehicles priced below fair market value, these entrepreneurs can target cars with strong resale potential, drawing on model-specific pricing data to make profitable decisions. Understanding how to spot undervalued vehicles—whether due to mileage, model, or production year—can be critical to securing a profitable purchase.

Additionally, in a competitive market where budget constraints drive many sellers to price vehicles aggressively, there are opportunities for car flippers to find quality vehicles at a discount. To capitalise on these opportunities, flippers require not only knowledge of current market trends and fair pricing but also insight into which models are more likely to attract buyers quickly. The data in this report can be used to guide these entrepreneurs in making strategic acquisitions, transforming undervalued vehicles into profit within Ireland's car market.

2 Chapter 2: Data

2.1 Collection

To gather data for this analysis, I developed a Python web scraping script using Selenium, an automation tool that enables web scraping. The script targeted DoneDeal, Ireland's largest online marketplace for used cars, capturing data daily over two defined periods: February 28, 2023, to April 5, 2023, and September 14, 2024, to September 27, 2024. These two windows serve as a snapshot to 2023 and 2024 respectively .

The data collected from DoneDeal included key vehicle information such as make, model, year, mileage, price, fuel type, focusing on cars listed under €7,000. This price threshold provides a comprehensive view of the affordable segment the Irish market which is often overlooked by other car pricing models such as the one found on Donedeal which does not provide valuations for cars more than 10 years old.

Additionally, the script was designed to track each listing's status over time, recording the date when a car was removed from DoneDeal, which is often a strong indicator of a sale. By capturing this information, the dataset enables analysis not only of the cars listed but also of those likely sold, offering insights into the pricing differences of those cars sold versus those that were not. This information can be utilised to construct a model that predicts the likelihood and duration required to sell a specific car at a designated price.

2.2 Data Cleaning

Data Cleaning was relatively straightforward, mainly removing cars with unlisted models, unlisted years and unlisted mileage. Since some cars were listed in Miles they were converted to Kilometers.

Cars with prices in pounds were also removed as those tended to mostly be cars registered in Northern Ireland which is outside the scope of this report. Cars with prices less than 500 were also removed as these tended to be non-car listings, such as advertising towing services or selling tires.

Cars models with less than 10 listings have also been removed, as it would be very hard to model these accurately due to the lack of data

2.3 Data Structure

The Python script has gathered the following attributes from each car listing

Attribute	Description
Link	Used to check if the listing has been removed.
Date Uploaded	Reflects the scraping date; useful for calculating listing duration and NCT validity.
Price	Price listed on the ad.
Make	Make of the car.
Model	Model of the car.
Trim	Name of the factory trim.
Trim Level	Equipment rating (base, medium, luxury)
Mileage (km)	Mileage in kilometers (converted from miles if necessary).
Fuel Type	Type of fuel: petrol, diesel, or hybrid.
Transmission	Automatic or manual transmission.
Body Type	Type of body (e.g., coupe, convertible, sedan, SUV).
Engine Size (Litres)	Size of the engine in liters.
Power	Horsepower of the car.
Acceleration (0-100)	Acceleration from 0 to 100 km/h.
Seats	Number of seats.
Doors	Number of doors.
Colour	Color of the car.
Country of Registration	Where the car was registered (only Ireland).
Previous Owners	Number of previous owners.
Road Tax	Road tax amount.
NCT Expiry	Month and year when the NCT expires.
Date Sold	Date when the listing was removed, suggesting a sale.
Sold Indicator	Whether the car was sold within the specified timeframe.

2.4 Exploratory Analysis

To start off this analysis, let's look at the price and mileage distribution of the top 6 most listed cars during the two years.

Firstly, the price distribution of the 6 most popular models in 2023 and 2024:

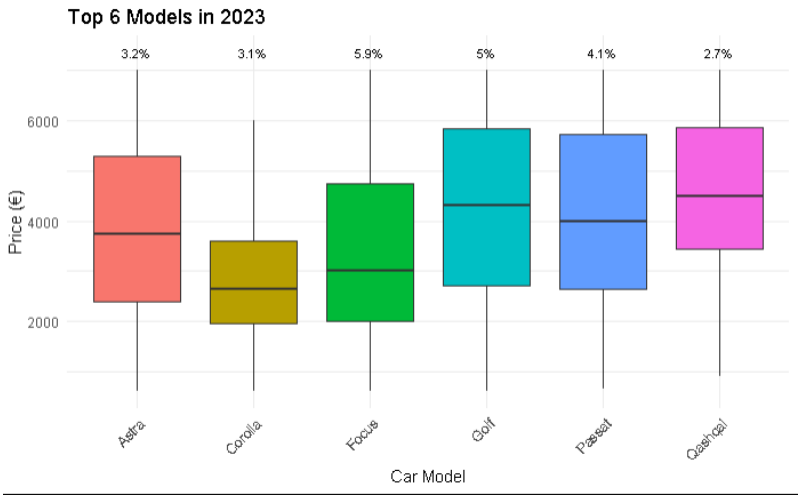


Figure 1: Price of Top 6 Models With Percent of Total Listings (2023)

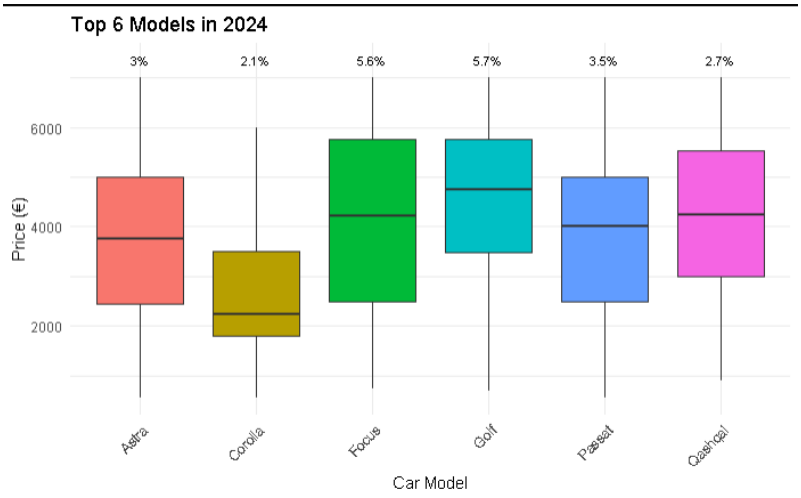


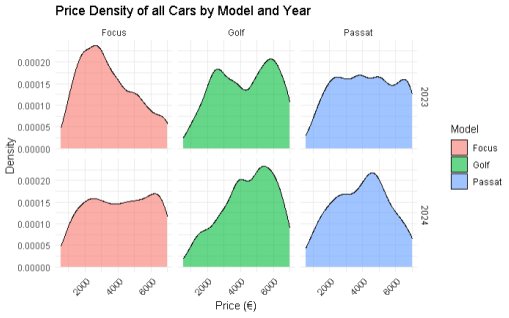
Figure 2: Price of Top 6 Models With Percent of Total Listings (2024)

The data shows that the Focus, Golf, and Passat were consistently the most popular cars listed in Ireland during 2023 and 2024. The table below shows the percentage of total listings each model represents for both years. These percentages indicate the share of listings occupied by each model.

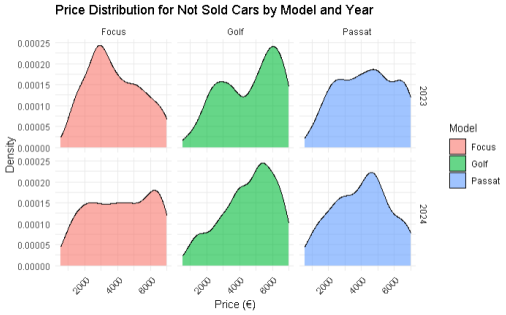
Model	2023 Percentage	2024 Percentage
Focus	5.9%	5.6%
Golf	5.0%	5.7%
Passat	4.1%	3.5%

Table 2: Top Car Models by Percentage of Listings in 2023 and 2024

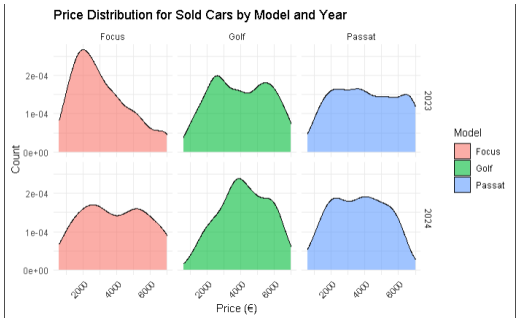
Now lets have a closer look at the price distribution of these models, divided by the year the car was listed on Donedeal.



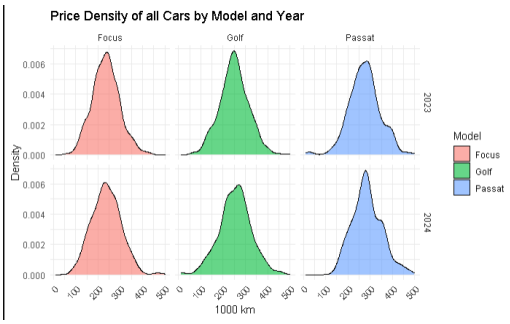
Price Distribution of All Cars (2023 and 2024)



Price Distribution of Not Sold Cars



Price Distribution of Sold Cars



Mileage Distribution of All Cars

Figure 3: Price and Mileage Distributions for Focus, Golf, and Passat

The four plots provide a comprehensive view of price distributions across car models (Focus, Golf, Passat) as well as mileage in 2023 and 2024 as well as highlighting the differences between sold and not sold cars.

The Price Density of All Cars plot for 2023 shows that Focus cars generally occupy lower price ranges, while Golf and Passat exhibit higher variance, with a more even distribution across all price points. In the 2024 plot, Passat and Golf distributions remain fairly consistent with those of 2023. However, there is a notable shift in the Focus distribution, which could be due to the introduction of a newer model of Focus that has depreciated below the €7,000 mark.

Comparing the price distributions of cars that were sold and those that were not sold, it can be observed that sold cars tend to be priced lower than cars that did not sell. This suggests that price may have a large effect on the likelihood of a car being sold.

The mileage of the Focus and Golf appears to be normally distributed around 250,000 km, while the mileage of the Passat has a mean value closer to 300,000 km. This suggests that the average mileage can vary between models

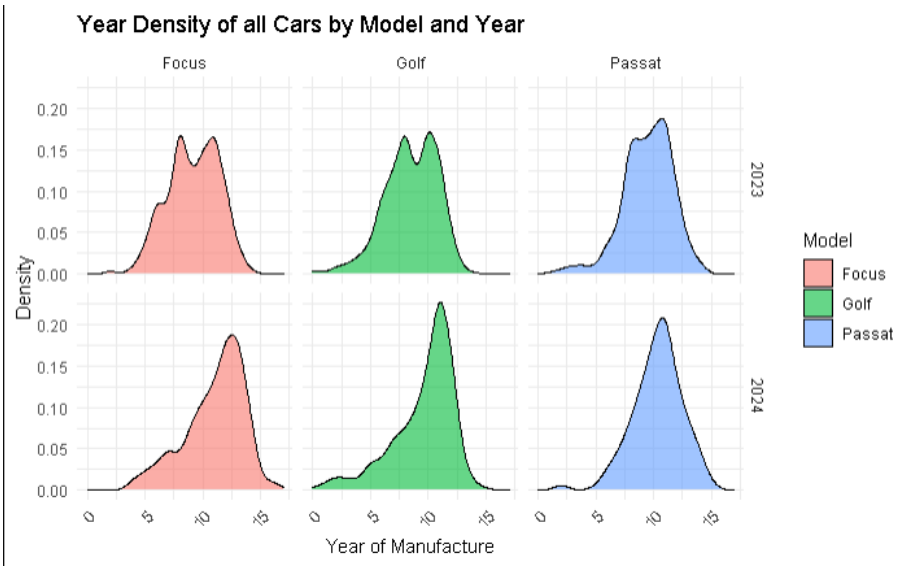


Figure 4: Year Density of Cars by Model and Year

The year density plot shows the distribution of manufacturing years for Focus, Golf, and Passat models in 2023 and 2024. Focus and Golf have two peaks, while Passat has a single peak, with distributions staying somewhat consistent across both years, with the most drastic change being the Focus distribution, it appears that the Focuses listed in 2024 were overall quite a bit newer than the ones listed in 2023

3 Chapter 3: Modelling

3.1 Linear Model

To begin, let's assess how well a **linear model** fits this data using the model:

$$\text{Price} \sim \text{Year} + \text{Model} + \text{Mileage}.$$

Initially, this model appears to fit quite well, with an adjusted R^2 of **0.6188** and significant p-values for both **Year** and **Mileage**, as well as mostly significant p-values for the majority of the car models.

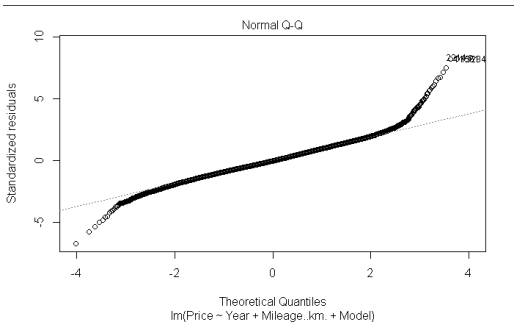


Figure 5: Residuals vs Fitted for the initial linear model.

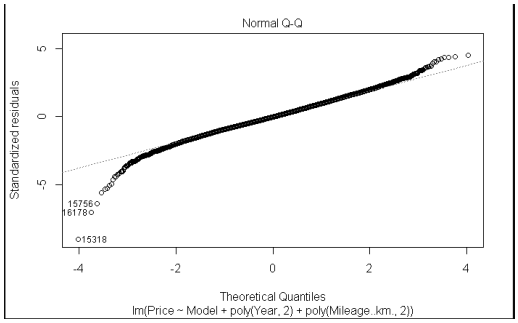


Figure 6: Polynomial transformation of the predictors in the linear model.

The **normal Q-Q plot** shows significant deviations from normality, particularly in the tails, where residuals at the extremes fall far above or below the expected values. This suggests that the model's residuals do not follow a normal distribution, which violates the assumptions of the linear model.

$$\text{Price} \sim \text{Year} + \text{Year}^2 + \text{Model} + \text{Mileage} + \text{Mileage}^2.$$

Adding **polynomial terms** helps somewhat with the normality of residuals but perhaps a better way capture the variability in car prices can be a **linear mixed-effects (LME) model**.

3.2 Linear Mixed Effects Model (LME)

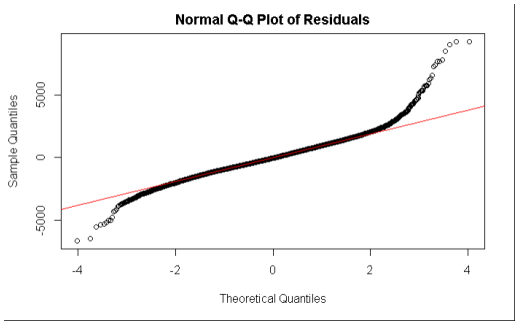
A **linear mixed-effects model** (LME) extends the standard linear model by incorporating both fixed and random effects. Fixed effects represent the overall impact of predictors (e.g., mileage, year), while random effects account for variability across different groups (e.g., car models).

An LME is suitable here as it allows for variability in car prices across different models. Unlike a standard linear model, which assumes a single pricing relationship for all models, the LME incorporates random effect allowing for the model to acknowledge that each car model may have a unique pricing relationship with factors like mileage and year, thereby capturing model-specific pricing patterns.

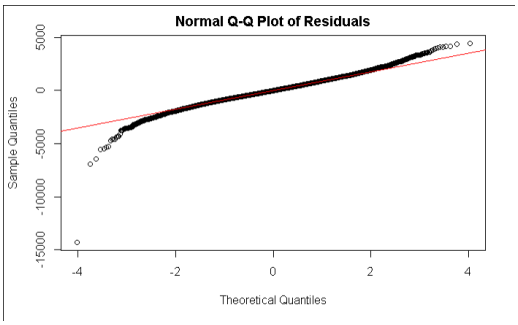
The initial LME model used is:

$$\text{Price} \sim \text{Mileage..km.} + \text{Year} + \text{Sold.Indicator}, \quad \text{random} = \sim 1 \,|\, \text{DatasetYear/Model}$$

Mileage and Year were chosen as they are strong predictors, and Sold.Indicator was included due to observed price differences between sold and unsold cars which was seen in the exploratory analysis. The random effects account for variability across dataset years (2023 or 2024) and models, allowing each model to have a unique intercept.



Q-Q Plot for LME



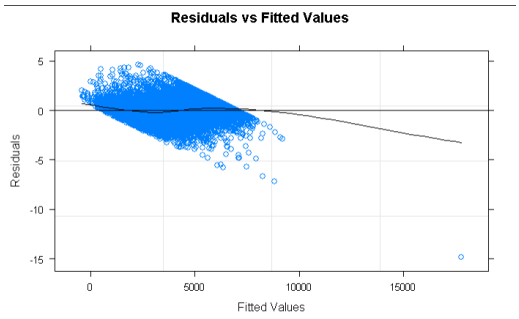
Q-Q Plot for Polynomial LME

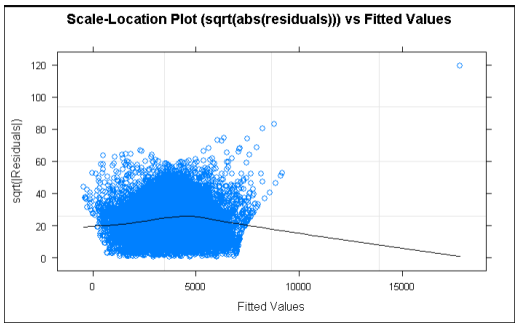
Figure 7: Comparison of Q-Q Plots for Polynomial and Non-Polynomial LME Residuals

Examining the Q-Q plot to test for the normal distribution of residuals reveals that this model also violates the assumption of normality, as such the addition of polynomial terms to Mileage and Year may be required.

$$\text{Price} \sim \text{poly}(\text{Mileage..km.}, 3) + \text{poly}(\text{Year}, 3) + \text{Sold.Indicator}, \quad \text{random} = \sim 1 \,|\, \text{DatasetYear/Model}$$

This has made the model more closely flow the normal distribution of residuals.





Residuals vs. Fitted Values **Scale-Location Plot**

Figure 8: Diagnostic Plots: Residuals vs. Fitted and Scale-Location

Although the Residuals vs. Fitted and Scale-Location plots are not flawless, they are reasonably acceptable, suggesting that the model remains sufficiently reliable for practical use.

4 Chapter 4: Observations

Lets look at the biggest changes from 2023 to 2024 data. Using an Anova test between 2023 and 2024 shows that only 5 cars went up in value, using a significance level of 0.05, which were are listed in the table below

Model	Mean Price 2023	Mean Price 2024	Price Difference	95% CI
Sportage	4161.57	5661.70	1500.13	(833.68, 2166.58)
S60	4243.24	5666.95	1423.71	(539.83, 2307.60)
Note	3364.98	4462.90	1097.91	(700.49, 1495.34)
Jetta	3613.78	4321.72	707.94	(294.95, 1120.93)
Focus	3421.85	4090.29	668.43	(512.05, 824.82)

Table 3: Price Differences, p-values, and Confidence Intervals for Selected Models (2023 vs. 2024)

While this may look quite conclusive, we must not forget about all the variables that can affect the price, for example, year, let's have a look at the density of Year between the two years below,

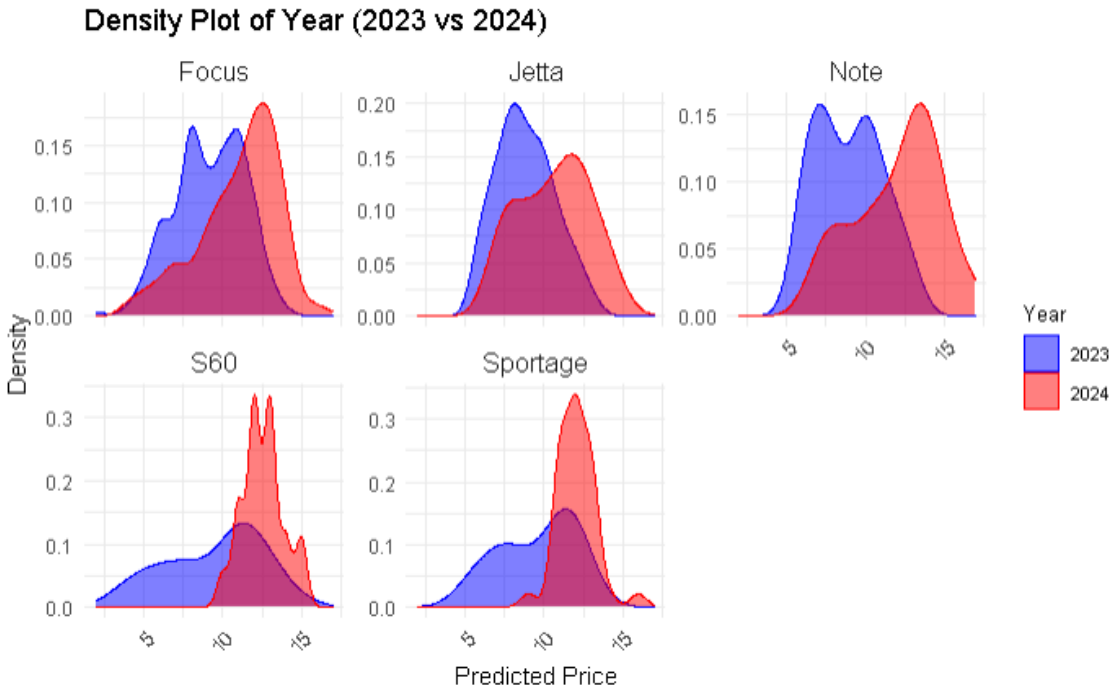


Figure 9: Manufacture Year Changes from 2023 to 2024

From this graph, this increase in price can potentially be attributed to the fact that in 2024, more newer cars were listed which usually correspond with higher prices, as such it is important to control these factors appropriately

Using the Linear Mixed Model outlined in Chapter 3, an analysis of the significant values with p-values less than 0.05 reveals that, after adjusting for year, mileage, and whether the car has been sold, no car shows a statistically significant increase in value.

Model	**Mean Price 2023**	**Mean Price 2024**	**Price Difference**	**95% CI**
Almera	1434.63	1213.74	-220.89	(-408.66, -33.12)
E-Class	4270.48	3922.94	-347.54	(-679.88, -15.22)
Jetta	4011.56	3658.71	-352.84	(-674.33, -31.35)
A3	4198.68	3809.92	-388.75	(-631.85, -145.66)
Insight	6059.17	5634.63	-424.53	(-745.36, -103.71)

Table 4: Price Differences and 95% Confidence Intervals for Selected Models (2023 vs. 2024)

Let us examine how the three most popular models have changed from 2023 to 2024, taking into account factors such as mileage, year of manufacture, and sale status, based on the results from our Linear Mixed Effects model

Model	**Mean Price 2023**	**Mean Price 2024**	**Price Difference**	**95% CI**
Focus	3829.14	3310.32	-518.81	[-648.99, -388.65]
Golf	4540.43	4018.93	-521.50	[-642.56, -400.44]
Passat	4331.91	3555.83	-776.08	[-915.78, -636.37]

Table 5: Price Differences and 95% Confidence Intervals for Selected Models (2023 vs. 2024)

When accounting for these factors, we can conclude that an identical Focus, valued at 3829 euros, has decreased in value by an amount between 648 euros and 388 euros, with 95% certainty. A similar conclusion can be drawn for the other models.

4.1 Predictions

Let us examine an example to illustrate how this model can be useful.

For instance, a 2010 Renault Megane is currently listed on DoneDeal (as of 11/11/24) for 1150 euros. To evaluate whether this price is reasonable, we can apply the model, taking into account that the vehicle has 152,000 kilometers on the odometer.

Putting this into our model we get the following result.

Model	Predicted Price	95% Confidence Interval
Megane	3110.85	[1374.08, 4847.61]

Table 6: Predicted Price and 95% Confidence Interval for Megane

Our model now suggests that based on the year and mileage of the car, its estimated value should be approximately 3110 euros. This implies that the car could represent a good deal and is well under fair market value.

However, it is important to note that purchasing a car is highly dependent on individual circumstances. Some vehicles may have mechanical issues that our model does not account for. Therefore, this pricing should only be considered as a rough estimate.

The 95% confidence interval is also quite wide, between 1347 and 4847, which can be attributed to the model's inability to account for the mechanical and physical condition of the vehicle.

4.2 Time to Sell Predictions

In certain situations, it's valuable to estimate the time it will take for a buyer to purchase a car. To model this, we can use a Negative Binomial Generalized Linear Model (GLM), which is often more robust than a Poisson GLM, particularly when dealing with overdispersed count data.

$$\begin{aligned} \text{days\_to\_sell} \sim\ & \text{poly}(\text{Price}, 2) + \text{poly}(\text{Mileage..km.}, 3) \\ & + \text{as.factor}(\text{Model}) + \text{Year} + \text{DatasetYear}, \\ & \text{Negative Binomial (log link)} \end{aligned} \tag{1}$$

Assuming we purchase the Renault Megane listed in the earlier example and price it at the suggested value of €3110, as predicted by our model, the expected average time to sell would be 3.6 days.

However, this model's accuracy is limited due to the reasons outlined in the shortfalls section. Furthermore, the GLM demonstrates a poor fit to the data. This should be used as a very rough estimate

Predicted Days to Sell	Lower Bound (95% CI)	Upper Bound (95% CI)
3.629081	3.020386	4.360447

Table 7: Predicted Days to Sell and 95% Prediction Interval

5 Chapter 5: Conclusion

5.1 Overview

In this project, I have developed a model aimed at accurately predicting the price of used cars and investigating the price differences between two years. I believe that the objectives have been successfully achieved. The model adheres well to the assumptions of a Generalized Linear Model (GLM) and provides a reliable estimate of the true market value of used cars.

5.2 Shortfalls

The biggest shortfall of this model lies in the data, namely while tweaking my web scraping program I accidentally overwrote the NCT.Expiry data for the 2024 dataset, which was a big mistake as in the 2023 data it proved to be a very good predictor of car price. This was not included as this report aimed more at predicting current 2024 car prices

The Time to sell is also a shortfall, this is due to the fact Donedeal auto reuploads cars to make them seem like they were recently put up for sale, ie a car could have been bumped up from 60 days ago and if the crawler would have found it then it would be labeled as new, and if it sold the next day the days to sell would be 1 instead of the correct 61. This error could be avoided if the crawler kept running for more days as eventually, Donedeal would stop re-uploading the ad

Another issue is that the model cannot capture the condition of the car, although NCT would have been a good indicator. Additionally, incorporating a web crawler to extract the description of the car listing and processing it through a Large Language Model to rate the condition on a scale from 1 to 10 could provide valuable insights into the Price.

I also believe this report would have benefitted from including additional graphs. However, I chose not to overload it with visualisations, as I plan to release the R script and a PowerBI report. These tools will provide more customisable visualisations and allow for deeper exploration of the data.

5.3 Further Study

Further research could explore increasing the price cutoff or eliminating it, allowing for a more comprehensive analysis of the entire car market.

Including location data, such as the county in which the advertisement was posted, which might enhance the accuracy of the model.

As previously mentioned, incorporating the description of the car could provide valuable insights. For instance, a car with a blown engine, which is usually mentioned in the description, would likely be priced significantly below market value, potentially leading to outliers in the model and resulting in less accurate predictions. A custom-trained Large Language Model, such as GPT, would be well-suited for analysing and extracting meaningful information and outputting a rating, say from 1-10 on how good it thinks the car is.