# Analys av begagnatpriser

## Inflytelserika faktorer och trender för Volkswagen-bilar


Martin Björkquist

EC Utbildning

Kunskapskontroll 2

202504

# Abstract

This project analyzed the prices of used cars in Sweden using data from Blocket to identify influential factors and develop a predictive model. By evaluating three linear regression models via 10-fold cross-validation, a model was selected based on factors such as age, mileage, horsepower, fuel type, transmission, seller type, car body type, and model group. The results indicate that age, mileage, and manual transmission have a significant negative impact on price, while higher horsepower and certain fuel types, like diesel, and in some cases, luxury models and minivans tend to increase it. Electric cars were found to be approximately 25% cheaper on average than gasoline cars, potentially reflecting market dynamics. The model explains about 94% ($R^2$) of the price variation in unseen data with a root mean squared error of approximately 36,000 SEK, suggesting a good ability to model the used car pricing in the Swedish market.

# Förkortningar och Begrepp

CV10: 10-faldig korsvalidering

EDA: Explorativ dataanalys

RMSE: Root Mean Square Error (medelkvadratfelets rot)

$R^2$: Förklaringsgrad (determinationskoefficient)

SCB: Statistiska Centralbyrån

VIF: Variance Inflation Factor (variationsinflationsfaktor)

Heteroskedasticitet: När variansen i residualerna inte är konstant över alla nivåer av prediktorerna, vilket kan påverka modellens tillförlitlighet för inferens.

Log-transformering: En matematisk transformation där den naturliga logaritmen appliceras på en variabel (t.ex. Försäljningspris) för att minska skevhet och förbättra linjäriteten i en regressionsmodell.

Multikollinearitet: När prediktorer i en regressionsmodell är starkt korrelerade med varandra, vilket kan leda till instabila skattningar av koefficienterna.

Outlier: En observation som avviker markant från övriga data och kan påverka modellens prestanda om den inte hanteras.

Prediktionsintervall: Ett intervall som anger osäkerheten kring en enskild prediktion. Det är bredare än ett konfidensintervall eftersom det inkluderar både osäkerhet i modellens uppskattning och variation i datan.

# Innehållsförteckning

# 1 Inledning

SCB-data ( SCB, 2025) visar att nyregistreringarna av personbilar i Sverige har förändrats avsevärt mellan 2006 och 2022. Medan bensin- och dieselbilar dominerade fram till 2018-2019, och etanolbilar hade en peak runt 2008, har elbilar och laddhybrider ökat kraftigt och stod för 54.5% av nyregistreringarna 2022. Den linjära trendlinjen för totalen indikerar en ökande utveckling av nyregistreringar över tid, med en genomsnittlig förändring på 9084 bilar per år, som framgår av Figur 1. Genom att analysera data från Blocket, en av Sveriges största marknadsplatser för begagnade bilar, syftar detta projekt till att utveckla en linjär regressionsmodell för att predicera försäljningspriser och identifiera de mest inflytelserika faktorerna, såsom bilens ålder, bränsletyp, och modellgrupp.

Projektet har två huvudsakliga mål:

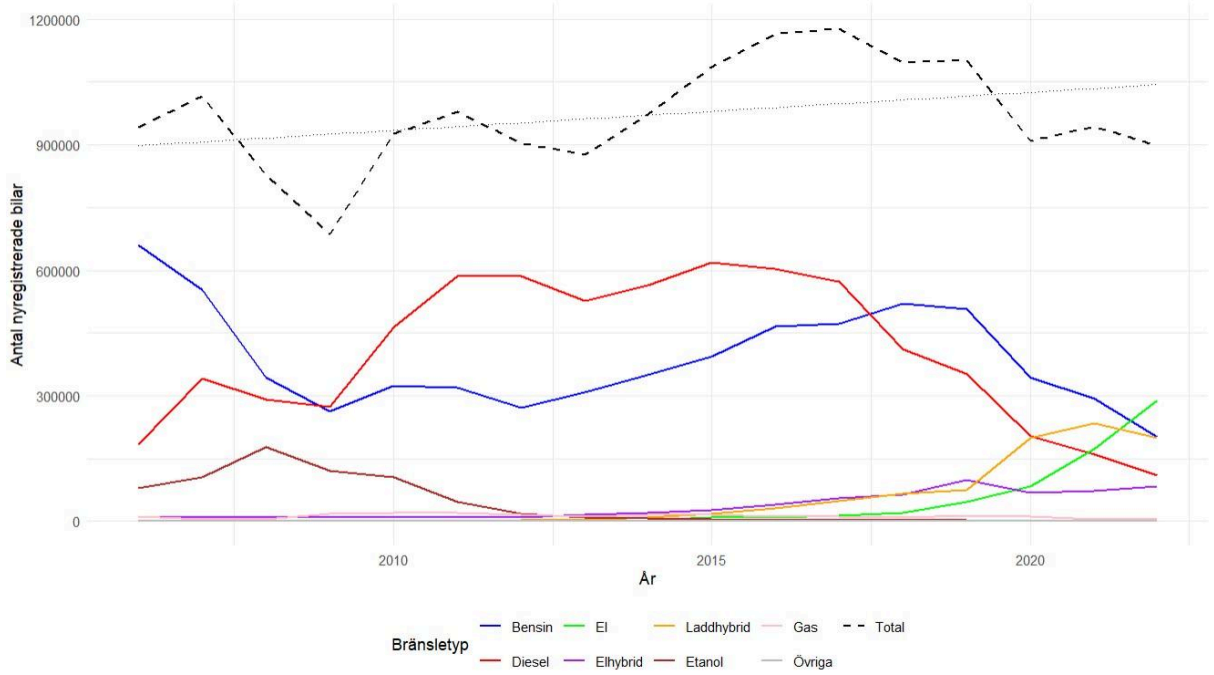
- att predicera försäljningspriser på osedd data med hög noggrannhet
- att utföra inferens för att förstå vilka faktorer som mest påverkar priset

Följande frågeställningar besvaras:

- Hur påverkar faktorer som bilens ålder, bränsletyp, och modellgrupp försäljningspriset?
- Kan vi utveckla en linjär regressionsmodell som både förklarar variationen i priset och predicerar priser på osedd data med hög noggrannhet?
- Hur hanteras extremvärden (outliers) för att förbättra modellens prestanda utan att riskera överanpassning?

Denna analys är relevant i ljuset av den pågående övergången till el- och hybridbilar, eftersom prisvariationen på begagnade bilar kan ge insikter om marknadens önskemål och framtida trender, vilket kan vara värdefullt för både konsumenter och bilbranschen.

Figur 1: Nyregistrerade personbilar per bränsletyp i Sverige, 2006–2022




*Bildtext: Diagram som visar nyregistrerade personbilar i Sverige uppdelat på bränsletyper mellan 2006 och 2022. Figuren illustrerar skiftet mot el- och hybridbilar och ger bakgrund till prisvariationer för begagnade Volkswagen-bilar i rapportens analys. Data hämtade från SCB:s dataset "Nyregistrerade personbilar efter län och kommun samt drivmedel. Månad 2006M01 - 2025M03" (SCB, 2025).*

## 1.1 Disposition

Rapporten är strukturerad enligt följande: Teoriavsnittet ger en bakgrund till linjär regression och begrepp som korsvalidering och outlier-hantering. Metodavsnittet beskriver datainsamlingen, datarensningen, den explorativa dataanalysen (EDA), modellbyggnadsprocessen, undersökningen av teoretiska antaganden och outlier-hanteringen mm. Resultatavsnittet presenterar modellernas prestanda och inferens. Diskussionen reflekterar över resultaten, metodens styrkor och svagheter, samt framtida förbättringsmöjligheter. Slutsatser sammanfattar de viktigaste fynden, och Appendix A innehåller kompletterande tabeller och figurer. Mot slutet finns också avsnitt med svar på teoretiska frågor och en självutvärdering.

# 2 Teori

## 2.1 Linjär regression

Linjär regression är en statistisk metod för att modellera ett linjärt samband mellan en beroende variabel (responsvariabel) och en eller flera oberoende variabler (prediktorer). Metoden används både för att förutsäga utfall och för att analysera hur olika faktorer påverkar ett resultat, till exempel för att identifiera vilka variabler – såsom ålder, körsträcka eller bränsletyp – som mest påverkar priset på begagnade bilar (James et al., 2021). I detta projekt är responsvariabeln det log-transformerade försäljningspriset (Log(Försäljningspris)), medan prediktorerna inkluderar kvantitativa variabler som ålder, körsträcka och hästkrafter, samt kategoriska variabler som bränsletyp, säljartyp, biltyp och modellgrupp. Modellen uttrycks matematiskt som:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

där ( Y ) är responsvariabeln, β0 är interceptet, βj är koefficienterna för prediktorerna Xj, och ϵ är en normalfördelad felterm med medelvärde 0 och konstant varians. Log-transformeringen av priset gör att koefficienterna kan tolkas som procentuella förändringar; exempelvis innebär en ökning av ålder med ett år en prisreduktion med β_Ålder% om övriga variabler hålls konstanta (James et al., 2021). Modellen är effektiv för att kvantifiera samband, förutsatt att dess antaganden om linjäritet, normalitet och konstant varians uppfylls.

## 2.2 Antaganden för linjär regression (Potentiella Problem)

Linjär regression kräver att flera antaganden uppfylls för att ge tillförlitliga resultat. Om dessa bryts kan modellens slutsatser bli missvisande (James et al., 2021). Potentiella problem är:

1. Linjäritet: Sambandet mellan responsvariabeln och prediktorerna ska vara linjärt, så att en förändring i en prediktor ger en proportionell förändring i responsvariabeln. Om sambandet är icke-linjärt fångar modellen inte det verkliga förhållandet.
2. Oberoende residualer: Residualerna (skillnaden mellan observerade och predikterade värden) ska vara oberoende och inte systematiskt korrelerade, t.ex. över tid. Korrelation mellan residualer kan leda till felaktiga koefficienter.
3. Homoskedasticitet: Residualerna ska ha konstant varians. Om variansen är icke-konstant (heteroskedasticitet) blir modellens precision och konfidensintervall opålitliga.
4. Normalfördelade residualer: Residualerna antas vara normalfördelade för att statistiska tester, såsom hypotesprövning av koefficienter, ska vara giltiga. Avvikelser kan påverka modellens slutsatser.
5. Outliers: Outliers är observationer som avviker kraftigt från det förväntade mönstret. De kan snedvrida modellens koefficienter och ge missvisande resultat.
6. High leverage-punkter: Observationer med extrema värden på prediktorerna (high leverage-punkter) är inflytelserika och kan påverka modellens lutning, även om de inte är outliers.
7. Multikollinearitet: När prediktorer är starkt korrelerade med varandra uppstår multikollinearitet, vilket gör koefficienterna instabila och svåra att tolka, eftersom deras individuella effekter blir otydliga.

## 2.3 Korsvalidering (CV10)

Korsvalidering utvärderar en modells prestanda på osedd data och säkerställer dess generaliserbarhet genom att minska risken för överanpassning (James et al., 2021). Vid 10-faldig korsvalidering (CV10) delas träningsdatan i 10 lika stora delar (folds). Modellen tränas på 9 folds och testas på den återstående, och processen upprepas 10 gånger så att varje fold används som testdata en gång. Prestandamåtten CV10 $R^2$ och CV10 RMSE beräknas som genomsnittet av $R^2$ respektive RMSE över iterationerna. CV10 $R^2$ visar hur mycket av variationen i responsvariabeln som modellen förklarar på osedd data, medan CV10 RMSE anger det genomsnittliga prediktionsfelet. Dessa mått är mer tillförlitliga än prestanda på träningsdatan, eftersom de simulerar modellens förmåga att generalisera (James et al., 2021).

## 2.4 Log-transformation

Log-transformering innebär att responsvariabeln transformeras med den naturliga logaritmen, i detta projekt tillämpat på försäljningspriset (Y = Log(Försäljningspris)) (James et al., 2021). Syftet är att hantera skevhet i prisdata, eftersom priser på begagnade bilar ofta är positivt skevade med en liten andel mycket höga värden. Log-transformering gör prisdistributionen mer symmetrisk och låter

koefficienterna tolkas som procentuella förändringar: en enhetsförändring i en prediktor Xj ändrar priset med cirka 100 x βj% (James et al., 2021).

## 2.5 $R^2$ (Determinationskoefficienten)

Determinationskoefficienten ($R^2$) mäter hur stor andel av variationen i responsvariabeln som modellen förklarar (James et al., 2021). Den beräknas som:

$R^2 = 1 - SSR/SST$

där SSR är summan av kvadrerade residualer och SST är den totala summan av kvadrater. $R^2$ ligger mellan 0 och 1, där ett värde nära 1 indikerar att modellen förklarar en stor del av variationen, medan ett värde nära 0 tyder på dålig passning.

## 2.6 RMSE (Root Mean Squared Error)

Root Mean Squared Error (RMSE) mäter det genomsnittliga prediktionsfelet genom att ta roten ur medelvärdet av kvadrerade residualer (James et al., 2021). Formeln är:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \dot{y}_i)^2}$$

där Yi är det observerade värdet, ŷi är det predikterade värdet och ( n ) är antalet observationer. I detta projekt är RMSE ca 36 000 SEK, vilket innebär att modellens prediktioner i genomsnitt avviker med 36 000 SEK från faktiska priser. Lägre RMSE indikerar bättre prediktionsnoggrannhet.

## 2.7 VIF (Variance Inflation Factor)

Variance Inflation Factor (VIF) mäter multikollinearitet genom att kvantifiera hur mycket variansen i en prediktors koefficient ökar på grund av korrelation med andra prediktorer (James et al., 2021). VIF för en prediktor Xj beräknas som:

$$VIF_j = \frac{1}{1 - R_j^2}$$

där $R^2j$ är determinationskoefficienten från en regression av Xj mot övriga prediktorer. Ett VIF-värde nära 1 indikerar låg multikollinearitet, medan värden över 5 eller 10 tyder på problematisk multikollinearitet, vilket kan ge instabila och svårtolkade koefficienter.

## 2.8 Outliers

Outliers är observationer som avviker kraftigt från datans mönster och kan snedvrida koefficienter samt försämra modellens prestanda. I detta projekt identifieras de med standardiserade residualer (värden > 3) och deras inflytande bedöms med Cook's distance, en vanlig metod (James et al., 2021).

## 2.9 Hypotesprövning

Hypotesprövning bedömer om en prediktor har en statistiskt signifikant effekt på responsvariabeln (James et al., 2021). För varje koefficient $\beta j$ testas nollhypotesen $H0: \beta j = 0$ (ingen effekt) mot alternativhypotesen $H1: \beta j \neq 0$. Ett p-värde under 0.05 leder till att H0 förkastas, vilket indikerar en signifikant effekt på responsvariabeln vid 5%-nivå.

## 2.10 Konfidensintervall

Ett 95%-konfidensintervall (CI) anger ett intervall där den sanna koefficienten $\beta j$ förväntas ligga med 95% sannolikhet, vilket visar osäkerheten i uppskattningen (James et al., 2021). För predikterade värden används konfidensintervall för att uppskatta medelvärdet av prediktionen, medan prediktionsintervall, som är bredare, uppskattar intervallet för en enskild observation, eftersom de inkluderar både modellens osäkerhet och datans variation.

# 3 Metod

## 3.1 Datainsamling

Datainsamlingen utfördes i två delar för att analysera begagnade bilpriser. Först hämtades data om nyregistrerade personbilar i Sverige (2006–2022) via ett API-anrop till SCB:s databas med ett R-skript. Detta möjliggjorde analys av trender i nyregistreringar per bränsletyp (t.ex. bensin, diesel, el), vilket gav kontext till prisvariationer mellan bränsletyper. Den andra delen bestod av en gruppinsamling av data från Blocket. Gruppen, bestående av Andreas Rasmusson, André Lindeberg, Ali Khalil, Karl Tengström, Gustav Jeansson, Svetlana Oshchepkova, Camilla Dahlman, Priyadarsini Panda, Oskar Amnér, Emil Nilsson, Lence Majzovska, Martin Blomqvist och undertecknad, samlade manuellt in data om 1 204 Volkswagen-personbilar (modellår 2000–2022) från geografiskt spridda regioner som Stockholm, Göteborg, Skåne, Norrbotten och andra län för att säkerställa ett representativt urval. Annonserna filtrerades för att exkludera yrkesfordon och leasing och sorterades efter senaste annonser (ej betalda placeringar). Insamlade variabler inkluderade försäljningspris, modellår, körsträcka, hästkrafter, bränsletyp, säljartyp (privat eller företag), biltyp (t.ex. SUV, kombi), växellåda, drivning, färg, modell, region och datum i trafik. Arbetet delades upp per region för att undvika dubbletter och standardiserades med en mall, vilket fungerade väl men ledde också till fel som inkonsekventa format, felplacerade värden och mellanslag. En viktig lärdom är att manuell datainsamling kräver noggrann rensning på grund av inkonsekvenser, till skillnad från strukturerade källor som SCB. Denna del av projektet gav värdefull erfarenhet av datakvalitet och vikten av tydliga instruktioner vid grupparbete.

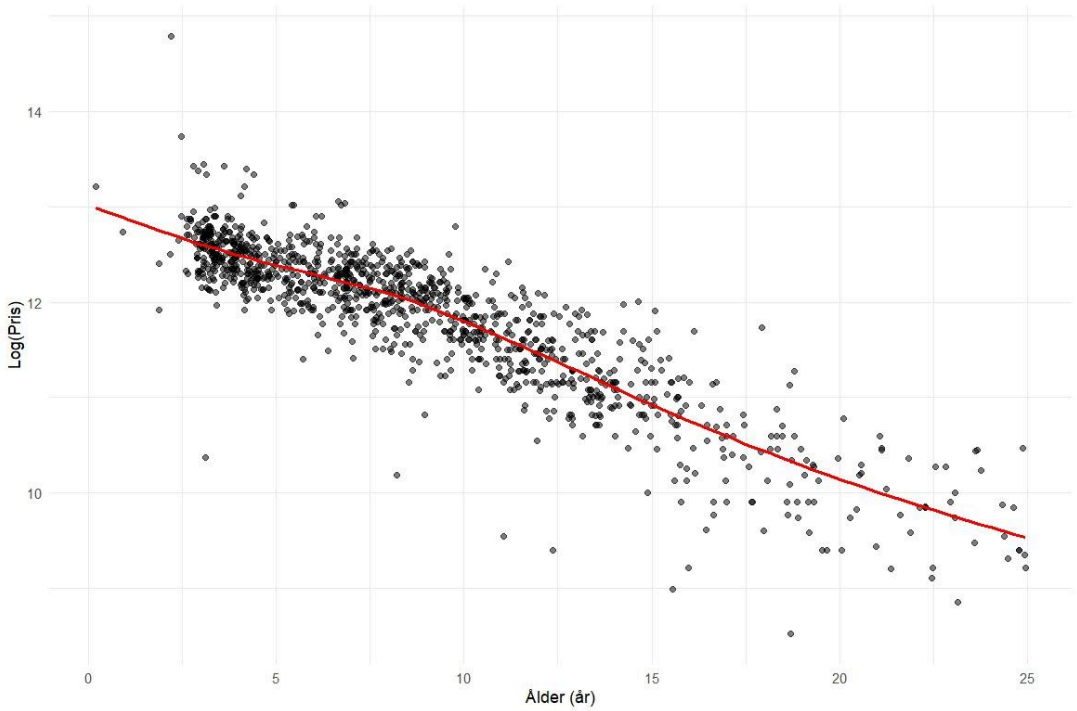
## 3.2 Datarensning

Datarensning utfördes för att säkerställa kvaliteten på Blocket-datan och förbereda den för analys. Datan laddades från Excel, där kolumner initialt var i textformat. Numeriska variabler, såsom försäljningspris, miltal, modellår och hästkrafter, rensades från mellanslag och kommatecken och omvandlades till numeriska värden. Kategoriska variabler, som bränsle, säljartyp och biltyp, konverterades till faktorer för modellering. Saknade värden hanterades genom att imputera medelvärdet för hästkrafter och ta bort rader med saknade värden i försäljningspris, miltal eller bränsle, vilket minskade datamängden från 1 204 till 1 192 observationer. Kategoriska variabler standardiserades genom att korrigera felstavningar (t.ex. "disel" till "diesel") och gruppera liknande kategorier (t.ex. "Cab" och "Sedan" till "Övriga" i biltyp) för att minska antalet nivåer, ett beslut som togs iterativt under modelleringen. Felaktiga värden korrigerades, såsom orimliga datum i Datum_i_trafik (t.ex. "20013-04-15" till "2013-04-15") och extrema miltal (t.ex. 249 764 mil till 24 976 mil). Observationer med orimliga datum (före 2000 eller efter 2025) eller modellår (efter 2022) togs bort. En ny variabel, Ålder, skapades som skillnaden mellan dagens datum och Datum_i_trafik för att fånga bilens användningstid. Ålder och modellår behölls temporärt för explorativ dataanalys (EDA), med avsikt att senare välja en för att undvika multikollinearitet. Två ytterligare variabler, Pris_per_Hästkraft (försäljningspris dividerat med hästkrafter) och Mil_per_År (miltal dividerat med Ålder), skapades för att analysera pris i förhållande till hästkrafter och körsträcka per år.

## 3.3 Explorativ dataanalys (EDA)

Explorativ dataanalys (EDA) genomfördes för att undersöka datans struktur, identifiera mönster och fatta informerade beslut inför modelleringen. Histogram av numeriska variabler som försäljningspris, miltal, hästkrafter, modellår, pris per hästkraft och mil per år (Appendix, Figur A1) visade att försäljningspris och miltal hade skev fördelning med lång högersvans, vilket motiverade log-transformering. Extrema värden, som en bil för 2 650 000 SEK, eller en bil med 402 610 mil identifierades som potentiella outliers. Hästkrafter och modellår hade mer förväntade fördelningar. Boxplot av försäljningspris mot kategoriska variabler som säljartyp, växellåda, biltyp, bränsle och region (Appendix, Figur A2–A5) visade prisskillnader, t.ex. högre medianpriser för företag, automatväxlade bilar, SUV:ar och elbilar, vilket motiverade deras inkludering som prediktorer. Spridningsdiagram mellan ålder och försäljningspris (Figur A10), samt ålder och log(försäljningspris) (Figur 2), bekräftade ett negativt samband som blev mer linjärt efter log-transformering, vilket stödde valet av log(försäljningspris) som responsvariabel. Korrelationsanalys av numeriska variabler (Appendix, Figur A6) visade starka samband, t.ex. mellan försäljningspris och ålder (-0.68), Försäljningspris och Hästkrafter (0.57) samt ålder och modellår (-0.98). Den höga korrelationen mellan ålder och modellår indikerade multikollinearitet, varför ålder valdes som prediktor för att bättre fånga bilens användningstid, och modellår exkluderades. Analys av bilmodeller (Appendix, Figur A7–A8) visade signifikanta prisskillnader mellan grupper (ANOVA, p-värde < 0.05), vilket motiverade inkluderingen av den skapade prediktorn modellgrupp.

Figur 2: Linjärt samband mellan ålder och log(försäljningspris)

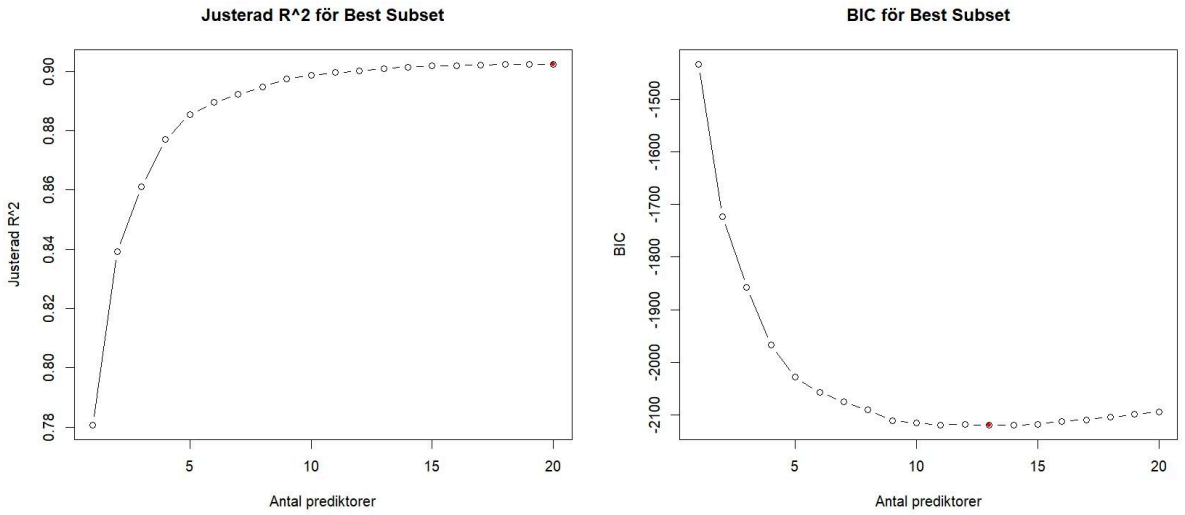



*Bildtext: Spridningsdiagram som visar sambandet mellan ålder och log(försäljningspris) för Volkswagen-bilar (modellår 2000–2022) baserat på Blocket-data. Efter log-transformering av försäljningspris framträder ett linjärt samband, till skillnad från det tidigare icke-linjära sambandet, vilket stödjer valet av log(försäljningspris) som responsvariabel. Priser över 500 000 SEK exkluderades för att fokusera på typiska bilar. Källa: Egen analys av Blocket-data.*

## 3.4 Modellbyggnad

Modellbyggnaden syftade till att utveckla en linjär regressionsmodell för att prediktera log(försäljningspris) baserat på bilattribut med fokus på inferens och prediktion (James et al., 2021). Datan delades slumpmässigt i ett träningsset (80 %, 954 observationer) och ett testset (20 %, 238 observationer) med reproducerbar uppdelning. Försäljningspris log-transformerades i R för att hantera skevhet, i linje med EDA-resultaten. Tre modeller testades: en basmodell (Model 1), en modell med variabelselektion (Model 2), och en modell med interaktionseffekter (Model 3). Prediktorer som drivning, märke och färg exkluderades. Exempelvis uteslöts färg på grund av subjektivitet och inkonsekvenser i kategorier (många varianter av samma grundfärg). Model 1 inkluderade alla relevanta prediktorer (ålder, miltal, hästkrafter, bränsle, säljartyp, biltyp, region, växellåda, modellgrupp). Ett rank-deficiensfel på grund av små nivåer i biltyp hanterades genom att gruppera kategorier till "Övriga" i datarensningen. Model 2 utvecklades med Best Subset Selection i R, där alla prediktorkombinationer utvärderades. Justerad $R^2$ och BIC användes för att välja en modell med 13 prediktorer tillhörande 8 huvudprediktorer (ålder, miltal, hästkrafter, bränsle, säljartyp, biltyp, växellåda, modellgrupp), baserat på lägsta BIC (Figur 3). Model 3 utökade Model 2 med en interaktionsterm mellan ålder och modellgrupp för att undersöka om prisets beroende av ålder varierade mellan modellgrupper. Alla modeller utvärderades med 10-faldig korsvalidering (CV10) i R, med prestandamått CV10 $R^2$ och RMSE, och koefficienternas signifikans bedömdes. Initialt valdes Model 3 för dess marginellt bättre prestanda och interaktionseffekter, men den förkastades senare på grund av höga VIF-värden (se Undersökning av teoretiska antaganden), och Model 2 valdes för vidare analys.

Figur 3: Justerad $R^2$ och BIC för Best Subset Selection av prediktorer




*Bildtext: Diagram som visar justerad $R^2$ (vänster) och BIC (höger) för Best Subset Selection av prediktorkombinationer på träningssetet. Den valda modellen med 13 prediktorer (kopplade till ålder, miltal, hästkrafter, bränsle, säljartyp, biltyp, växellåda, modellgrupp) markeras vid lägsta BIC, vilket optimerar prediktionen av log(försäljningspris) i Model 2. Källa: Egen analys i R.*

## 3.5 Undersökning av teoretiska antaganden

Undersökningen av teoretiska antaganden syftade till att utvärdera om linjär regressions antaganden uppfylldes för Model 3 och sedan Model 2, samt identifiera potentiella problem (James et al., 2021). Multikollinearitet bedömdes med Variance Inflation Factor (VIF) i R. Höga VIF-värden för Model 3, särskilt för interaktionstermen ålder:modellgrupp, indikerade problematisk multikollinearitet och efter att ha övervägt att slå ihop fler kategorier i modellgrupp valdes istället Model 2, vars VIF-värden var acceptabla och skillnaden ändå så liten mot Model 3's CV10 $R^2$ och RMSE (Figur A13). För Model 2 undersöktes icke-linjäritet med en Residuals vs Fitted-plot (Appendix, Figur A9), som bekräftade linjäritet tack vare log-transformeringen av responsvariabeln. Oberoende observationer verifierades med Durbin-Watson-testet, som inte visade autokorrelation (Figur A12). Heteroskedasticitet analyserades med en Scale-Location-plot (Appendix, Figur A9), som antydde viss icke-konstant varians, vilket noterades för tolkningen. Normalfördelning av residualerna bedömdes med en Q-Q-plot (Appendix, Figur A9), som visade approximativ normalfördelning, med mindre avvikelser i svansarna som ett antal extrempunkter bidrog till. Dessa extrempunkter identifierades som outliers med standardiserade residualer, där värden över 3 markerades för vidare hantering (se Outlier-hantering). High-leverage-punkter noterades med hat-värden, men hanterades inte ytterligare (Figur A15). Inflytelserika observationer undersöktes med Cook's distance, där inga punkter hade oproportionerligt inflytande (Figur A11).

## 3.6 Outlier-hantering

Outlier-hantering syftade till att identifiera och hantera extremvärden som kunde påverka modellens prestanda och tillförlitlighet (James et al., 2021). Baserat på standardiserade residualer > 3 från undersökningen av teoretiska antaganden identifierades 15 outliers (Figur A14). En okulär analys genomfördes för att bedöma deras karaktär, och fyra observationer bedömdes som särskilt avvikande: en rallybil med priset 2 650 000 SEK, två bilar med körsträckor över 40 000 mil, och en nyare SUV med orimligt lågt pris (Figur A16). Dessa fyra observationer (index 285, 626, 734, 847) togs bort från träningssetet, vilket gav ett reducerat dataset med 950 observationer. För att motivera borttagningen utvärderades Model 2:s prestanda med 10-faldig korsvalidering (CV10) med och utan de borttagna observationerna. En modell utan alla 15 outliers testades också, men borttagning av de fyra prioriterades för att balansera robusthet och generaliserbarhet. CV10-jämförelsen (se Resultat och diskussion) stödde beslutet att använda det reducerade datasetet i den slutgiltiga modellen.

## 3.7 Utvärdering mot osedd data

Utvärdering mot osedd data genomfördes för att bedöma den slutgiltiga modellens generaliseringsförmåga (James et al., 2021). Model 2 tränades om på det reducerade träningssetet (950 observationer) efter borttagning av fyra outliers, med samma prediktorer som tidigare (ålder, miltal, hästkrafter, bränsle, säljartyp, biltyp, växellåda, modellgrupp). Prediktioner gjordes på testsetet

(238 observationer) i R, och prestandan mättes med $R^2$ och RMSE på log-skala (log(försäljningspris)) samt RMSE i SEK efter återtransformering med exponentiering. Koefficienternas signifikans och modellens passning bedömdes för att säkerställa att prediktorerna fortfarande bidrog meningsfullt till inferens. Prestandaskillnader mellan tränings- och testdata noterades för vidare diskussion (se Resultat och diskussion).

## 3.8 Hypotesprövning och konfidensintervall

Hypotesprövning och konfidensintervall syftade till att bedöma prediktorernas signifikans och osäkerhet i den slutgiltiga Model 2 samt utvärdera prediktionernas tillförlitlighet (James et al., 2021). Hypotesprövning genomfördes i R, där p-värden för prediktorer analyserades för att identifiera statistiskt signifikanta bilattribut ($p < 0.05$). Konfidensintervall (95 %) för koefficienterna beräknades för att mäta osäkerheten. För att bedöma prediktionernas tillförlitlighet valdes de första fem observationerna från testdatan. Konfidensintervall för medelvärdet av predikterade log-priser och prediktionsintervall för individuella log-priser (95 %) beräknades och exponentierades till SEK för att kunna tolkas. Resultaten (se Resultat och diskussion) användes för att dra slutsatser om prediktorer och prediktiv förmåga till rapporten.

# 4 Resultat och Diskussion

Analysen resulterade i en linjär regressionsmodell (Model 2) för att predicera log(försäljningspris) för begagnade Volkswagen-bilar. Jämförelsen mellan Model 1, 2 och 3 visade att Model 3 hade marginellt bättre CV10 $R^2$-prestanda ($R^2$: 0.9032 vs 0.9031) men Model 3 förkastades på grund av multikollinearitet (se Appendix, Tabell A1). Explorativ dataanalys (EDA) blev vägledande bl.a. genom att identifiera starka korrelationer, som mellan försäljningspris och ålder, och behovet av log-transformering för att hantera skevhet (James et al., 2021). Modellen utvärderades i flera steg, och resultaten kopplades bl.a. till marknadstrender för att ge insikter om bilattributens påverkan på priset.

## 4.1 Modellprestanda på träningsdata

Modellens prestanda utvärderades med 10-faldig korsvalidering (CV10) på träningssetet under olika outlier-scenarier (Tabell 1). Efter borttagning av fyra extrema observationer (en rallybil, två bilar med hög körsträcka, en billig SUV) förbättrades prestandan markant:

- CV10 $R^2$: Ökade från 0.9031 (SD: 0.0394) till 0.9206 (SD: 0.0252).
- CV10 RMSE (log-skala): Minskade från 0.2478 (SD: 0.0564) till 0.2234 (SD: 0.0396).

Borttagning av alla 15 outliers gav ännu bättre resultat ($R^2$: 0.9359, RMSE: 0.1886), men detta valdes bort för att undvika överanpassning.

Tabell 1: CV10-resultat för Model 2 under olika outlier-scenarier

| **Scenario** | **CV10 $R^2$ (SD)** | **CV10 RMSE, log-skala (SD)** |
|---|---|---|
| Ingen borttagning | 0.9031 (0.0394) | 0.2478 (0.0564) |
| 4 outliers borttagna | 0.9206 (0.0252) | 0.2234 (0.0396) |
| 15 outliers borttagna | 0.9359 (0.0121) | 0.1886 (0.0192) |

*Bildtext: Tabell som visar 10-faldig korsvalidering (CV10) resultat för Model 2 på träningssetet under tre outlier-scenarier: ingen borttagning, borttagning av fyra extrema observationer, och borttagning av alla 15 outliers. Prestandamått ($R^2$ och RMSE i log-skala, med standardavvikelser) illustrerar hur borttagning av fyra outliers förbättrar prediktionen av log(försäljningspris) med mindre risk för överanpassning. Källa: Egen analys i R.*

Tabell 2. Prestandamått för Model 2 på testdatan.

| **Mått** | **Värde** |
|---|---|
| $R^2$ | 0.9378 |
| RMSE (log-skala) | 0.1948 |
| RMSE (SEK) | 35 852 |

*Bildtext: Tabell som visar prestandamått för Model 2 på testdata (238 observationer), inklusive $R^2$, RMSE i log-skala, och RMSE i SEK. Resultaten ($R^2$: 0.9378, RMSE: 35 852 SEK) indikerar stark generaliseringsförmåga för prediktion av log(försäljningspris), med rimliga prediktionsfel. Källa: Egen analys i R.*

## 4.2 Utvärdering på testdata

Den slutgiltiga modellen uppvisade stark generaliseringsförmåga på testdatan (238 observationer), med följande prestandamått (Tabell 2):

- $R^2$: 0.9378, vilket indikerar att modellen förklarar 93.8 % av variationen i log(försäljningspris).
- RMSE (log-skala): 0.1948.
- RMSE (SEK): 35 852 SEK, vilket visar på rimliga prediktionsfel i ursprunglig skala.

Prestandan var oväntat bättre än på träningssetet, vilket kan bero på färre extrema observationer i testdatan eller skillnader i datadistribution. Detta diskuteras vidare nedan.

## 4.3 Prediktorernas påverkan

Hypotesprövning visade att flera prediktorer var statistiskt signifikanta ($p < 0.05$), och koefficienterna (Tabell 3) gav insikter om bilattributens effekter på priset.

Viktiga resultat inkluderar:

- Ålder: En ökning med ett år minskar priset med cirka 8.9 %.
- Bränsle:
    - Elbilar är i genomsnitt 25.0 % billigare än bensinbilar. Detta resultat kan verka motsägelsefullt eftersom rådatan visar att elbilar har ett högre medelpris (293 041 SEK) än bensinbilar (148 779 SEK), och Figur A5 bekräftar att elbilar generellt är dyrare över regioner. Skillnaden beror på att elbilar i urvalet är betydligt nyare (medelålder 3.77 år vs. 10.3 år för bensinbilar), har färre mil (7 028 vs. 12 876), fler hästkrafter (220 vs. 135), och oftare tillhör dyrare modellgrupper som SUV:ar (56.5 % vs. 17.2 % för bensinbilar). Modellen justerar för dessa effekter.
    - Dieselbilar är 13.8 % dyrare.
- Modellgrupp: Lyxmodeller och minibussar är 19.3 % respektive 30.4 % dyrare än familjebilar.

Konfidensintervall (95 %) bekräftade tydliga effekter. Till exempel varierar elbilar från 19.7 % till 30.0 % billigare än bensinbilar (konfidensintervall: -0.3557 till -0.2201). För de första fem observationerna i testdatan beräknades konfidens- och prediktionsintervall och för den första observationen var:

- Konfidensintervall (medelpris, SEK): 156 364–177 714.
- Prediktionsintervall (individuellt pris, SEK): 107 021–259 651, vilket återspeglar större osäkerhet för enskilda prediktioner.

Tabell 3: Signifikanta prediktorer och koefficienter i Model 2

| **Prediktor** | **Koefficient** | **Tolkning (prisförändring i %)** |
|---|---|---|
| Ålder | -0.0929 | -8.9 % per år |
| Miltal | -0.00002735 | -0.003 % per mil |
| Hästkrafter | 0.004125 | +0.4 % per hk |
| Bränslediesel | 0.1291 | +13.8 % (jämfört med bensin) |
| Bränsleel | -0.2879 | -25.0 % (jämfört med bensin) |
| Bränslemiljöbränsle/hybrid | -0.09211 | -8.8 % (jämfört med bensin) |
| SäljarePrivat | -0.1228 | -11.6 % (jämfört med företag) |
| BiltypFamiljebuss | 0.1288 | +13.7 % (jämfört med Övriga) |
| Växellådamanuell | -0.1305 | -12.2 % (jämfört med automat) |
| ModellgruppKompaktbilar | 0.05753 | +5.9 % (jämfört med Familjebilar) |
| ModellgruppLyxmodeller | 0.1767 | +19.3 % (jämfört med Familjebilar) |
| ModellgruppMinibussar | 0.2651 | +30.4 % (jämfört med Familjebilar) |
| ModellgruppÖvriga | 0.2467 | +28.0 % (jämfört med Familjebilar) |
| ModellgruppSUV:ar | 0.1686 | +18.4 % (jämfört med Familjebilar) |

*Bildtext: Tabell som visar statistiskt signifikanta prediktorer (p-värde < 0.05) och deras koefficienter i Model 2. Tolkningen i procent anger hur varje prediktor påverkar priset, t.ex. en åldersökning med ett år minskar priset med 8.9 %. Källa: Egen analys i R.*

## 4.4 Diskussion

Resultaten kopplade till SCB:s dataset 'Nyregistrerade personbilar efter län och kommun samt drivmedel. Månad 2006M01 - 2025M03', visar att el- och laddhybridbilar utgjorde 54.5 % av nyregistreringarna 2022 (SCB, 2025), en trend som sannolikt förstärkts 2023–2025 på grund av förbättrad batteriteknik. Trots detta är begagnade elbilar 25.0 % billigare än bensinbilar, vilket kan bero på:

- Snabb teknisk utveckling som gör äldre elbilar mindre attraktiva.
- Osäkerheter kring batterilivslängd och laddinfrastruktur.
- Höga elpriser i Sverige 2021–2022, som kan ha minskat efterfrågan på elbilar.

Dieselbilar är 13.8 % dyrare, vilket kan reflektera deras popularitet före elbilstrenden (t.ex. 2010–2015, då diesel dominerade nyregistreringarna enligt SCB-datan). Ålder och modellgrupp (t.ex. lyxmodeller, minibussar) påverkar priset starkt, vilket gör modellen till ett bra verktyg för att predicera priser och vägleda köpare och säljare i begagnatmarknaden. Ytterligare resultat finns i Appendix (se Tabell A1 och Figur A11–A20).

# 5 Slutsatser

Denna rapport har besvarat tre centrala frågor om prediktion av begagnatpriser för Volkswagen-bilar.

Den första frågan – hur faktorer som bilens ålder, bränsletyp, och modellgrupp påverkar försäljningspriset – visade att ålder minskar priset med cirka 8.9 % per år, och lyxmodeller och minibussar är betydligt dyrare (se Tabell 3). Elbilar är 25.0 % billigare än bensinbilar när man justerar för andra variabler, trots att rådatan visar ett högre medelpris för elbilar (293 041 SEK vs. 148 779 SEK för bensinbilar). Detta beror på att elbilar är nyare (medelålder 3.77 år vs. 10.3 år) och oftare tillhör dyrare modellgrupper som SUV:ar (56.5 % vs. 17.2 %), vilket modellen justerar för. Dieselbilars högre priser (13.8 %) kan troligen kopplas till deras popularitet före elbilstrenden (SCB, 2025). Dessa resultat ger insikter om vilka bilattribut som påverkar priset på begagnatmarknaden.

Den andra frågan – om en linjär regressionsmodell kan förklara prisvariation och predicera priser på osedd data – besvarades genom utvecklingen av Model 2. Modellen förklarar 92.1 % av prisvariationen på träningsdatan ($R^2$=0.9206, Tabell 1) och uppvisade hög noggrannhet på testdatan med 93.8 % ($R^2$=0.9378 och RMSE=0.1948 i log-skala) (Tabell 2). Prediktionsintervall för enskilda observationer (Figur A20) visade praktisk användbarhet för köpare och säljare som med modellen har ett bra verktyg för att uppskatta priser med hög tillförlitlighet.

Den tredje frågan – hur extremvärden (outliers) hanteras för att förbättra modellens prestanda utan att riskera överanpassning – besvarades genom en noggrann analys av datan. Extremvärden identifierades med statistiska metoder som standardiserade residualer och Cook's distance (se Figur A11 och A14) och en tabell med de 15 mest extrema observationerna (Figur A16) användes för okulär kontroll, där exempelvis rallybilen och bilar med extremt hög körsträcka granskades. Beslutet att behålla eller ta bort dessa outliers baserades på deras påverkan på modellens prediktionsförmåga. Genom att selektivt ta bort ett fåtal extremvärden förbättrades Model 2:s prestanda, med en $R^2$ på 0.9206 och en RMSE på 0.2234 i log-skala (se Tabell 1), utan att öka risken för överanpassning alltför mycket. Detta visar att en balanserad hantering av extremvärden är avgörande för att skapa en generaliserbar modell för prediktion av begagnatpriser.  Projektet har dock begränsningar:

- Heteroskedasticitet i residualerna kan påverka inferensens tillförlitlighet.
- Analysen är begränsad till Volkswagen, vilket gör generaliserbarheten till andra märken osäker.
- Modellen fångar inte regionala variationer i pris, vilket kan påverka tolkningen av bränsleeffekter (se Figur A5).

Framtida forskning kan undersöka bredare dataset eller andra bilmärken, och inkludera region som prediktor för att bättre förstå prisvariationer. Sammantaget visar projektet hur statistisk modellering ger värdefulla insikter om prisbildning och understryker vikten av att förstå bilattribut och marknadstrender. Ytterligare detaljer om modellens utveckling och diagnostik finns i Appendix A (se Tabell A1 och Figur A9–A20).

# 6 Teoretiska frågor

1. Kolla på följande video: https://www.youtube.com/watch?v=X9_ISJ0YpGw&t=290s , beskriv kortfattat vad en Quantile-Quantile (QQ) plot är.

*QQ plot används oftast för att visa om ett dataset är normalfördelad genom att jämföra dess kvantiler mot de teoretiska kvantilerna i en perfekt normalfördelning. Om data är normalfördelad så kommer punkterna ligga i en någorlunda rät linje (45 grader). Ifall punkterna avviker så kan det visa på att datan inte är normalfördelad och mönstret på avvikelsen berättar om hur, t.ex. skevhet och tjocka eller tunna svansar.*

2. Din kollega Karin frågar dig följande: ”Jag har hört att i Maskininlärning så är fokus på prediktioner medan man i statistisk regressionsanalys kan göra såväl prediktioner som statistisk inferens. Vad menas med det, kan du ge några exempel?” Vad svarar du Karin?

*I maskininlärning så är man i huvudsak intresserad av att hitta en modell som kan förutsäga så bra som möjligt och det är själva prediktionens förmåga som är det viktiga, inte varför den gör den prediktionen eller vad som påverkar den. T.ex. vid spamklassificering där det viktiga är hur effektiv modellen är. I statistisk inferens så är man då dessutom intresserad av hur de olika oberoende variablerna påverkar den beroende variabeln och att hitta samband som kan hjälpa oss dra slutsatser. T.ex. för huspriser där faktorer som läge, storlek, antal rum mm. påverkar men vi vill veta hur mycket och ifall det finns samband, synergieffekter osv.*

3. Vad är skillnaden på ”konfidensintervall” och ”prediktionsintervall” för predikterade värden?

*Konfidensintervall uppskattar intervallet för medelvärdet av prediktionen, medan prediktionsintervall uppskattar intervallet för en enskild observation. Prediktionsintervallet är bredare eftersom det inkluderar både osäkerheten i modellen och den naturliga variationen i datan.*

4. Den multipla linjära regressionsmodellen kan skrivas som:

$Y = \beta 0 + \beta 1x1 + \beta 1x2 + \ldots + \beta pxp + \varepsilon$ .

Hur tolkas beta parametrarna?

*β0, interceptet, är värdet på den beroende variabeln (Y) när alla oberoende variabler x är noll. Övriga β parametrar är koefficienter för vardera variabler x och visar förändringen i Y då x ökar en enhet, när alla andra oberoende variabler hålls konstanta.*

5. Din kollega Nils frågar dig följande: ”Stämmer det att man i statistisk regressionsmodellering inte behöver använda träning, validering och test set om man nyttjar mått såsom BIC? Vad är logiken bakom detta?” Vad svarar du Hassan?

*Ja, men inte om primära syftet är att modellen ska hitta den bästa prediktionen, då är det fortfarande bra med att dela upp med ett testset också för att få en validering på osedd data. Men skulle det till övervägande del skulle vara för att hitta en enkel och modell så kan det funka bra eftersom BIC straffar komplexa modeller med alltför många parametrar som annars leder till överanpassning.*

6. Förklara algoritmen nedan för ”Best subset selection”

**Algorithm 6.1** *Best subset selection*

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly $k$ predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest $R^2$.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using using the prediction error on a validation set, $C_p$ (AIC), BIC, or adjusted $R^2$. Or use the cross-validation method.

*Best subset selection hittar bästa möjliga kombinationen av oberoende variabler för att förutspå beroende variabeln Y.*

*1. I först läget M0 gissas Y enbart genom ett genomsnitt på äldre värden. Inga variabler används.*

*2. Sedan testas alla variabler. Först en variabel i taget för att hitta den enskilt bästa för att prediktera Y. Sedan testas kombinationer av två variabler för att hitta vilka två som tillsammans ger bäst prediktion av Y. Därefter fortsätter testerna med de bästa tre, fyra osv. tills alla variabler körts igenom.*

*Algoritmen väljer ut den bästa kombinationen för varje antal kombinerade variabler (1st, 2 st, 3st osv.) med hjälp av lägst möjliga RSS-värde eller största R^2 värde.*

3. I sista steget ska en modell med bra balans väljas ibland dessa, en enkel modell är att föredra. Valet görs med hjälp av Crossvalidation prediction error eller BIC, AIC och adjusted R^2 som straffar komplexa modeller.

7. Ett citat från statistikern George Box är: “All models are wrong, some are useful.” Förklara vad som menas med det citatet.

*Ingen modell är felfri men vissa hjälper oss med tillräckligt bra förutsägelser och insikter samt att dra slutsatser om vår data.*

# 7 Självutvärdering

1. Vad tycker du har varit roligast i kunskapskontrollen?

*Att det kändes som ett "riktigt" uppdrag på ett sätt med olika datakällor och utmaningar med att data sätts samman ifrån flera håll eller personer. Tvätt och EDA kändes som att jobba :)*
*Det har också varit intressant och kanske även roligt att iterativt gå fram och tillbaka mellan datarensning, EDA, modellbyggandet och till slut ändå komma i mål, många turer blir det, men man lär sig sjukt mycket med allt som behöves efterforskas eller läsas på om mellan dessa steg. Tyckte det var lärorikt också att jobba i grupp vid insamlingen, både för samarbetets skull men också för att sedan ta hand om den "smutsiga" data som ändå blir vid insamling på det viset. Envist ville jag göra allt i R dessutom så det har blivit extremt mycket kod och har kanske t.o.m. börjat gilla R litegrann.*

2. Hur har du hanterat utmaningar? Vilka lärdomar tar du med dig till framtida kurser?

*Denna gången har jag med mig att tackla rapportskrivandet annorlunda. Provar att skriva på svenska denna gången för jag vet att jag på min LIA kommer använda mig mest av det och har försökt använda ett mer akademiskt språkbruk ihop med kortare och mer koncisa stycken. Det blir som en omskolning för mig som kämpat för det motsatta lång tid, att skriva allt enklare texter med vardagliga uttryck, uppdelade med mycket mellanrum osv för omsorgspersonal. Jag tycker att min "berättelse" tappas litegrann när jag skriver såhär men kan förstå att det behöver vara så och jag fortsätter öva, det blir dock fortfarande mycket text. Annars är jag nöjd med min projektplanering och upplägg av arbetsgången i de två senaste kunskapskontrollerna så utmaningen har kanske mer varit att tiden innan varit kort med att hinna förkovra sig till nivån man skulle önska. Men som sagt, man lär sig nog mer via att komma igång med projekten ändå.*

3. Vilket betyg anser du att du ska ha och varför?

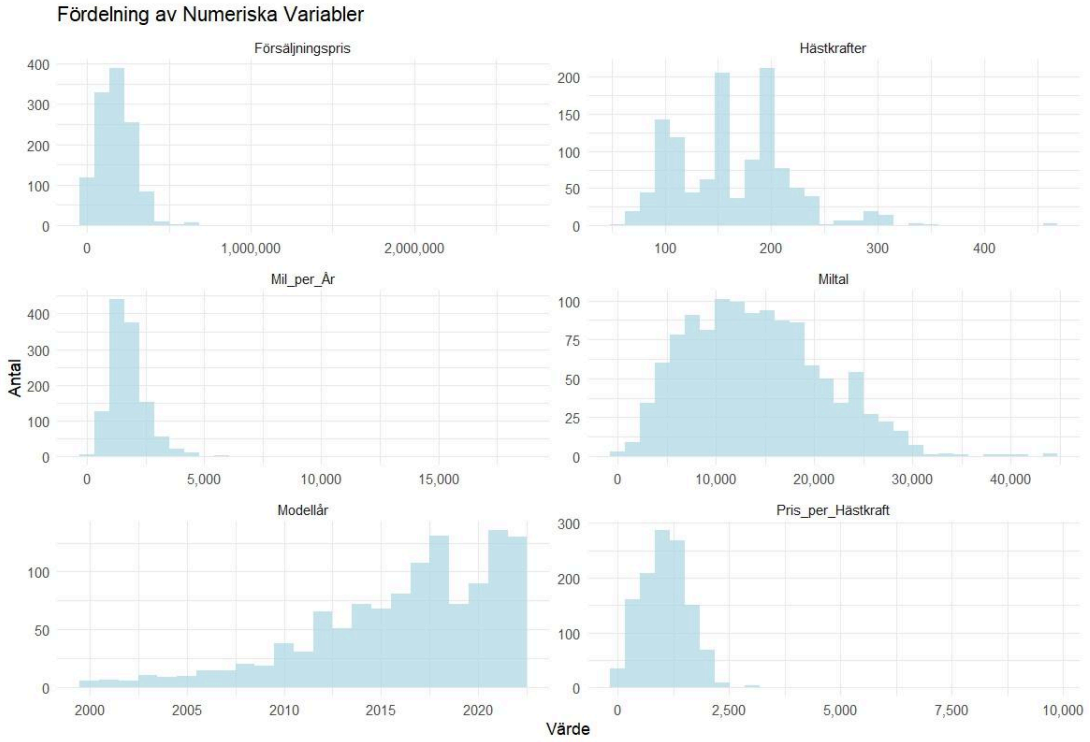
*Jag siktar inte på någon särskild betygsnivå men har försökt nå upp till eller utföra VG-nivåns uppgifter, mest för att få de kunskaperna som då krävdes i uppgifterna och i denna kunskapskontrollen var de ju roligare än G-nivåns. Jag hoppas såklart att jag blir godkänd och skulle vara nöjd med det, medveten om att ribban har höjts med denna kursen.*

4. Något du vill lyfta till Antonio?

*Nej*

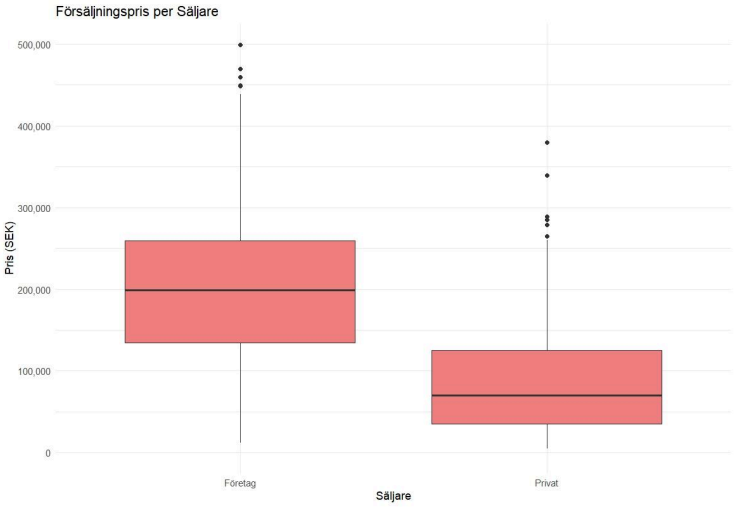
# Appendix A

Figur A1: Histogram för numeriska nyckelvariabler



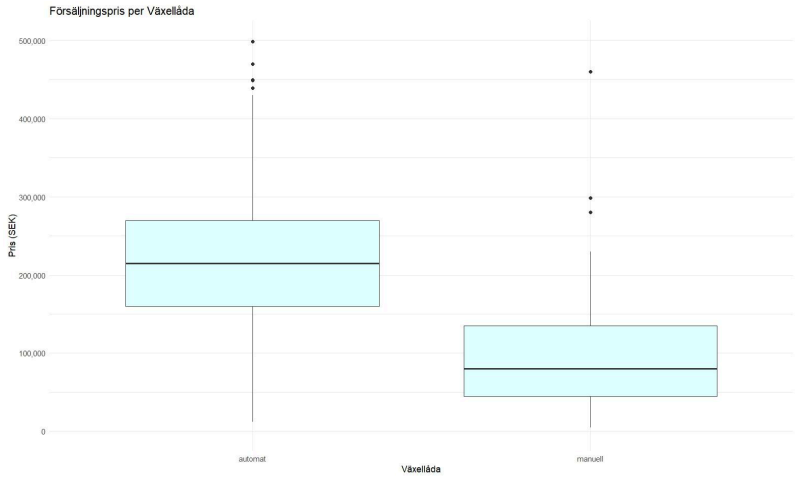

*Bildtext: Kombinerat barchart med sex histogram som visar fördelningarna av försäljningspris (SEK), miltal (mil), hästkrafter (hk), modellår, pris per hästkraft (SEK/hk), och mil per år. Diagrammet illustrerar variablernas spridning och eventuella snedheter. Källa: Egen analys i R.*

Figur A2: Försäljningspris per säljartyp



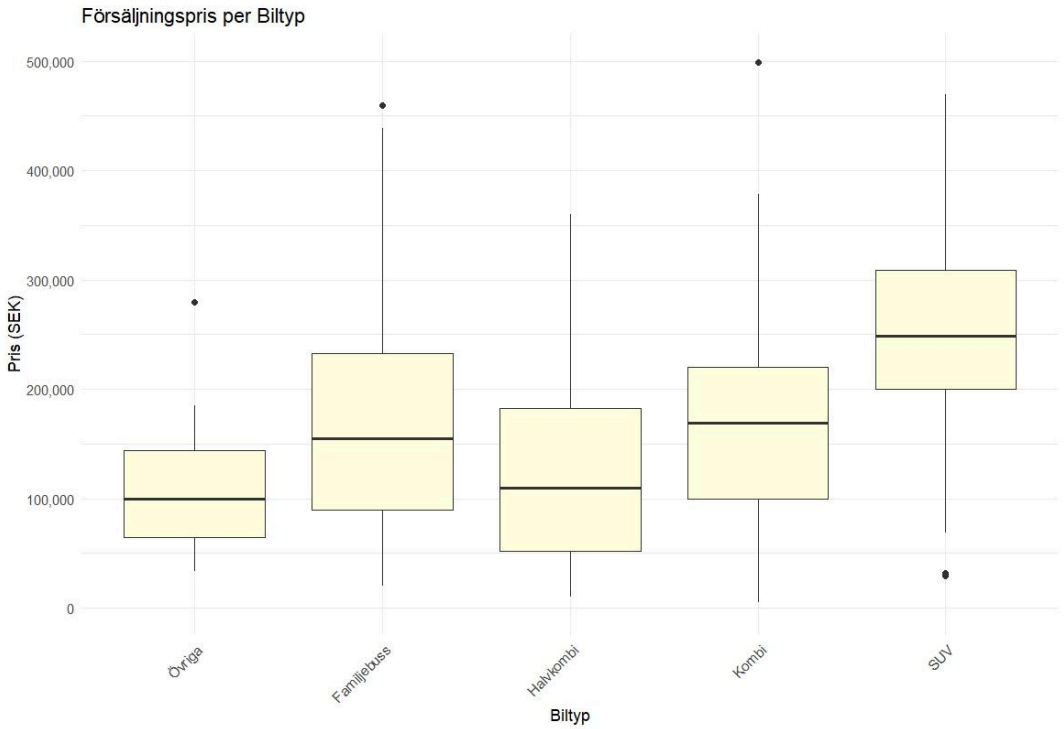

*Bildtext: Boxplot som visar fördelningen av försäljningspris (SEK) uppdelat på säljartyp (privat eller företag). Diagrammet visar medianpriser, spridning, och eventuella outliers, vilket indikerar att företag ofta har högre priser. Källa: Egen analys i R.*

Figur A3: Försäljningspris per växellådstyp



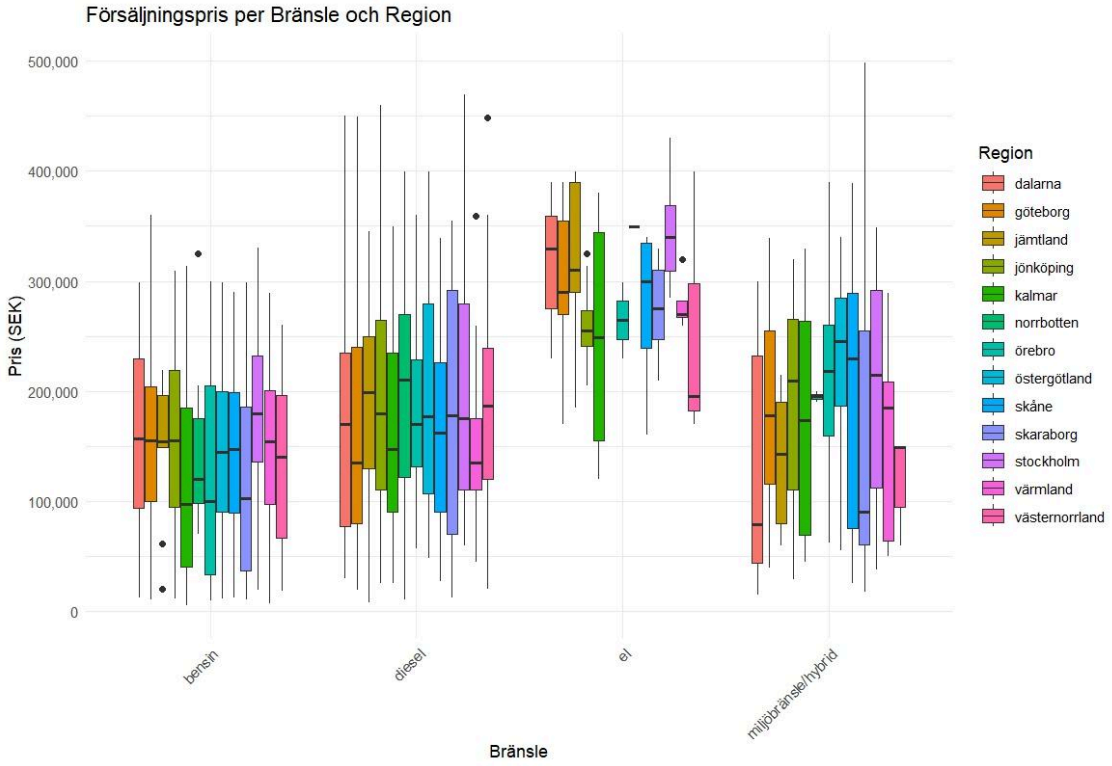

*Bildtext: Boxplot som visar försäljningspris (SEK) uppdelat på växellådstyp (manuell eller automat). Diagrammet illustrerar att automatiska bilar generellt har högre medianpriser. Källa: Egen analys i R.*

Figur A4: Försäljningspris per biltyp



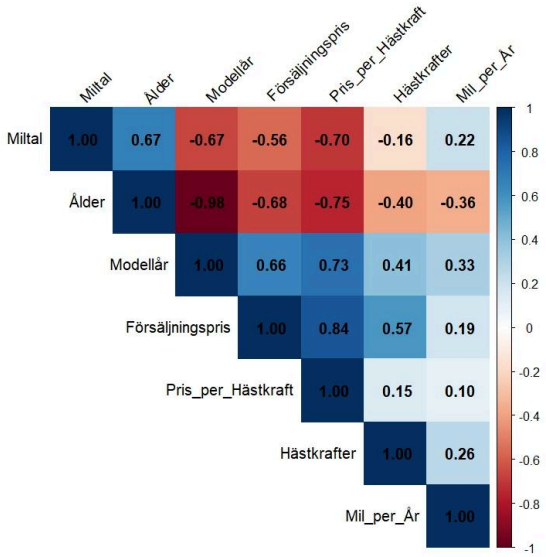

*Bildtext: Boxplot som visar fördelningen av försäljningspris (SEK) uppdelat på biltyp (t.ex. familjebuss, SUV). Diagrammet visar att vissa biltyper, som SUV:ar, har högre medianpriser, medan andra har större spridning. Källa: Egen analys i R.*

Figur A5: Försäljningspris per bränsletyp och region



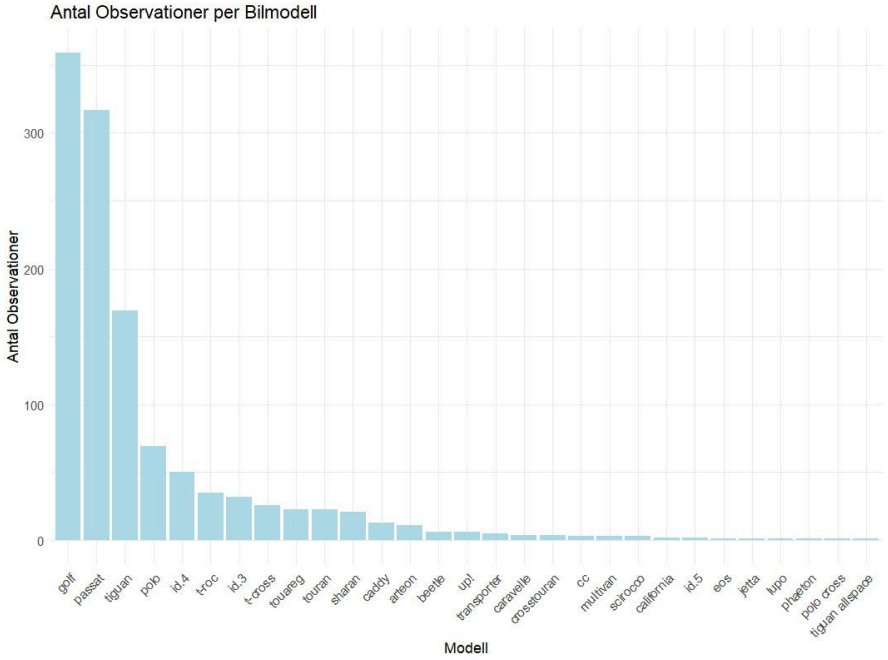

*Bildtext: Boxplotar som visar försäljningspris (SEK) uppdelat på bränsletyp (bensin, diesel, el och miljöbränsle/hybrid) och region. Källa: Egen analys i R.*

Figur A6: Korrelationsmatris för numeriska variabler



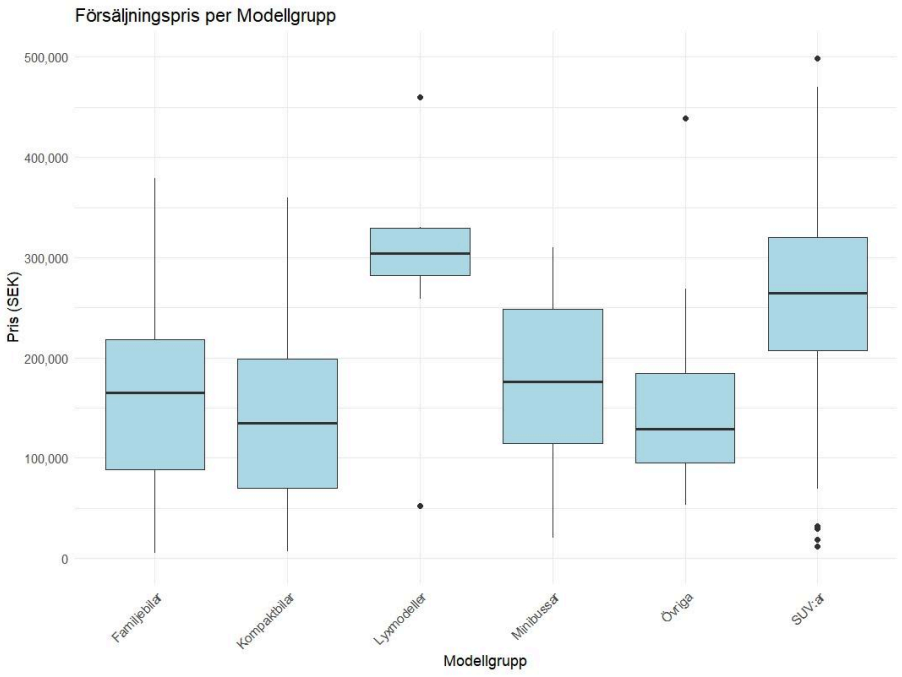

*Bildtext: Korrelationsmatris som visar sambanden mellan numeriska variabler (t.ex. försäljningspris, miltal, hästkrafter, ålder). Starka korrelationer, t.ex. mellan ålder och pris, indikerar viktiga prediktorer för modellering av log(försäljningspris). Källa: Egen analys i R.*

Figur A7: Antal observationer per Volkswagen-modell



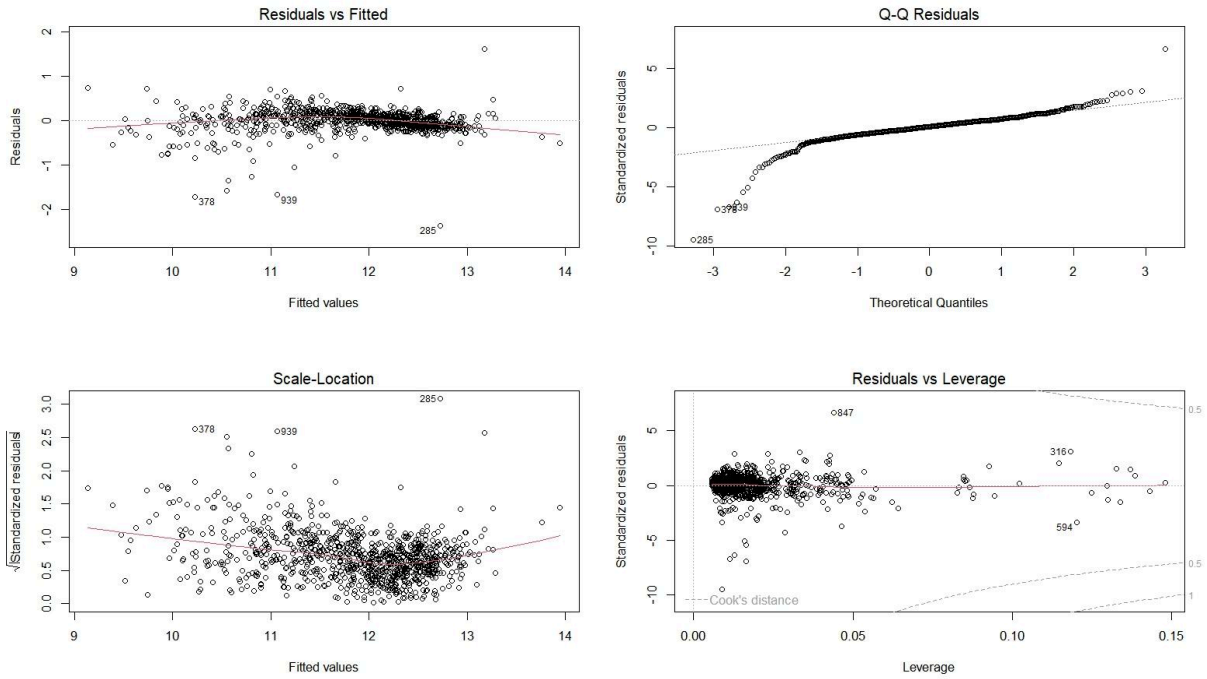

*Bildtext: Stapeldiagram som visar antalet observationer per bilmodell för Volkswagen-bilar i datamängden. Diagrammet illustrerar att vissa modeller (t.ex. Golf) är överrepresenterade, vilket kan påverka modellens generaliseringsförmåga. Källa: Egen analys i R.*

Figur A8: Försäljningspris per modellgrupp



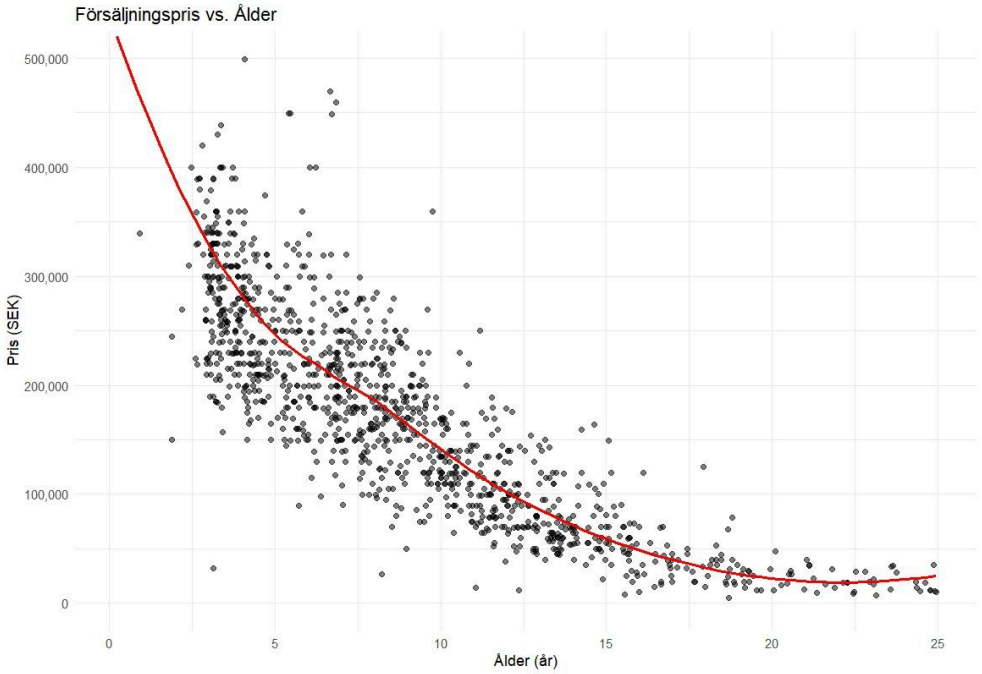

*Bildtext: Boxplot som visar försäljningspris (SEK) uppdelat på modellgrupp (t.ex. kompakta bilar, lyxmodeller). Diagrammet visar att lyxmodeller och SUV:ar har högre medianpriser. Källa: Egen analys i R.*

Figur A9: Diagnostiska plottar för Model 2




*Bildtext: Fyra diagnostiska plottar för Model 2, inklusive Residuals vs Fitted, Q-Q plot, Scale-Location, och Residuals vs Leverage. Källa: Egen analys i R.*

Figur A10: Försäljningspris kontra ålder före log-transformering




*Bildtext: Spridningsdiagram som visar försäljningspris (SEK) kontra bilens ålder (år) före log-transformering. Diagrammet visar ett icke-linjärt samband, vilket motiverar log-transformering av försäljningspris för bättre modellpassning. Källa: Egen analys i R.*

Figur A11: Inflytelserika observationer baserat på Cook's distance

```
Inflytelserika observationer (Cook's distance > 0.5):
> print(influential)
named integer(0)
>
```

*Bildtext: Utskrift som visar observationer med Cook's distance över 0.5 för Model 2, vilket skulle indikera potentiellt inflytelserika observationer. Källa: Egen analys i R.*

Figur A12: Durbin-Watson-test för Model 2

```
> # Testa för korrelerade residualer (Durbin-Watson test)
> dw_test <- dwtest(model2)
> cat("Durbin-Watson-test:\nDW =", dw_test$statistic, "p-value =", dw_test$p.value, "\n")
Durbin-Watson-test:
DW = 1.954296 p-value = 0.2409048
>
```

*Bildtext: Utskrift som visar resultatet av Durbin-Watson-testet för Model 2:s residualer. Testet indikerar låg autokorrelation, vilket stödjer att residualerna är oberoende och att modellen är lämplig för prediktion av log(försäljningspris). Källa: Egen analys i R.*

Figur A13: VIF-värden för prediktorer i Model 2

```
VIF-värden för Model 2:
> print(vif_values)
                GVIF Df GVIF^(1/(2*Df))
Ålder       3.032206  1        1.741323
Miltal      2.503764  1        1.582329
Hästkrafter 1.905784  1        1.380501
Bränsle     2.255959  3        1.145219
Säljare     1.373495  1        1.171962
Biltyp     18.382831  4        1.438969
Växellåda   1.803832  1        1.343068
Modellgrupp 19.105209  5        1.343121
>
```

*Bildtext: Utskrift som visar Variance Inflation Factor (VIF)-värden för prediktorerna i Model 2. Låga VIF-värden (under 5) indikerar att multikollinearitet inte är ett problem, vilket stödjer modellens stabilitet för prediktion av log(försäljningspris). Källa: Egen analys i R.*

Figur A14: Lista över outliers i datamängden

```
> # Identifiera potentiella outliers baserat på standardiserade residualer
> std_res <- rstandard(model2)
> outliers <- which(abs(std_res) > 3)
> cat("Outliers (observationer med |std residual| > 3):", outliers, "\n")
Outliers (observationer med |std residual| > 3): 217 263 285 316 361 378 549 594 626 734 834 847 886 909 939
>
```

*Bildtext: Lista som visar observationer identifierade som outliers baserat på statistiska kriterier (t.ex. standardiserade residualer). Dessa outliers analyserades för att avgöra om de skulle tas bort för att förbättra prediktionen av log(försäljningspris). Källa: Egen analys i R.*

Figur A15: Observationer med hög leverage i Model 2

```
> # Identifiera observationer med hög leverage
> leverage <- hatvalues(model2)
> p <- length(coef(model2))
> n <- nrow(train_data)
> high_leverage <- which(leverage > 2 * p / n)
> cat("High leverage-punkter:", high_leverage, "\n")
High leverage-punkter: 57 69 79 89 91 96 98 102 111 135 136 137 140 177 184 185 203 208 244 289 298 316 320 327 329 333 339
366 393 404 414 419 435 449 457 473 493 498 526 527 528 535 547 552 560 569 575 582 591 594 604 625 628 630 638 644 659 667
672 684 695 698 703 727 734 740 742 744 758 790 797 802 811 814 819 847 857 885 903 910 911 924 925 951 952
>
```

*Bildtext: Utskrift som visar observationer med hög leverage i Model 2, baserat på leverage-värden. Källa: Egen analys i R.*

Figur A16: Tabell över de 15 mest extrema outliers

```
Outliers i träningsdata:
> print(train_data[outliers_to_check, ])
# A tibble: 15 × 19
   Försäljningspris Säljare Bränsle Växellåda Miltal Modellår Biltyp     Drivning Hästkrafter Färg  Datum_i_trafik Märke
              <dbl> <fct>   <fct>   <fct>      <dbl>    <dbl> <fct>      <fct>          <dbl> <fct> <date>         <fct>
 1            12000 Privat  bensin  manuell    20490     2006 Halvkombi  2wd              102 silv… 2005-08-25     volk…
 2             9900 Privat  bensin  manuell    23641     2004 Kombi      2wd              105 brun  2003-12-12     volk…
 3            31900 Företag diesel  automat     7860     2022 SUV        4wd              151 grå   2022-03-04     volk…
 4           459900 Företag diesel  manuell    14460     2018 Familjebu… 4wd              150 vit   2018-06-21     volk…
 5            26500 Privat  diesel  automat    24600     2000 Halvkombi  2wd               62 silv… 2017-02-01     volk…
 6             5000 Privat  bensin  manuell    27213     2007 Kombi      2wd              150 silv… 2006-08-14     volk…
 7             9000 Privat  bensin  manuell    27000     2003 Kombi      2wd              150 grå   2002-11-07     volk…
 8            52500 Privat  diesel  automat    29357     2007 Övriga     4wd              225 svart 2013-09-16     volk…
 9             8000 Privat  diesel  manuell    28741     2010 Kombi      4wd              141 silv… 2009-10-05     volk…
10            19999 Privat  diesel  manuell    40261     2009 Familjebu… 2wd              175 blå   2009-07-10     volk…
11            19500 Privat  diesel  manuell    44103     2001 Kombi      2wd               90 vit   2000-12-21     volk…
12          2650000 Privat  bensin  manuell       86     2018 Halvkombi  4wd              272 vit   2023-02-01     volk…
13            10000 Privat  diesel  manuell    23862     2009 Kombi      4wd              105 blå   2009-04-27     volk…
14            14000 Privat  bensin  manuell    24021     2001 Halvkombi  2wd               75 brun  2014-04-02     volk…
15            11999 Privat  diesel  automat    25660     2010 Kombi      2wd              141 silv… 2012-12-11     volk…
```

*Bildtext: Tabell som listar de 15 mest extrema outliers i datamängden, inklusive variabler som försäljningspris, miltal, och ålder. Tabellen användes för okulär kontroll ifall outliers (t.ex. rallybilen med högt pris eller bilar med hög körsträcka) bör tas bort. Källa: Egen analys i R.*

Figur A17: Sammanfattning av Model 2

```
> summary(final_model)

Call:
lm(formula = formula_model2, data = train_data_4)

Residuals:
     Min       1Q   Median       3Q      Max
-1.71927 -0.09996  0.00976  0.12324  0.71072

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                1.244e+01  6.058e-02 205.304  < 2e-16 ***
Ålder                     -9.294e-02  2.571e-03 -36.151  < 2e-16 ***
Miltal                    -2.735e-05  1.671e-06 -16.362  < 2e-16 ***
Hästkrafter                4.125e-03  1.929e-04  21.379  < 2e-16 ***
Bränslediesel              1.291e-01  2.020e-02   6.391 2.59e-10 ***
Bränsleel                 -2.879e-01  3.456e-02  -8.331 2.85e-16 ***
Bränslemiljöbränsle/hybrid -9.211e-02  2.631e-02  -3.501 0.000486 ***
SäljarePrivat             -1.228e-01  2.047e-02  -5.999 2.85e-09 ***
BiltypFamiljebuss          1.288e-01  5.855e-02   2.200 0.028022 *
BiltypHalvkombi           -5.100e-02  4.662e-02  -1.094 0.274306
BiltypKombi               -8.544e-02  4.357e-02  -1.961 0.050164 .
BiltypSUV                 -1.028e-01  5.310e-02  -1.937 0.053093 .
Växellådamanuell          -1.305e-01  2.074e-02  -6.291 4.84e-10 ***
ModellgruppKompaktbilar    5.753e-02  2.580e-02   2.230 0.025987 *
ModellgruppLyxmodeller     1.767e-01  6.493e-02   2.721 0.006628 **
ModellgruppMinibussar      2.651e-01  5.923e-02   4.476 8.57e-06 ***
ModellgruppÖvriga          2.467e-01  8.021e-02   3.076 0.002162 **
ModellgruppSUV:ar          1.686e-01  3.592e-02   4.694 3.09e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2234 on 931 degrees of freedom
Multiple R-squared:   0.92,     Adjusted R-squared:  0.9185
F-statistic: 629.4 on 17 and 931 DF,  p-value: < 2.2e-16
```

*Bildtext: Utskrift som visar sammanfattningen av Model 2 i R, inklusive koefficienter, p-värden, $R^2$, och residual standardfel. Källa: Egen analys i R.*

Figur A18: Signifikanta prediktorer i Model 2

```
Signifikanta prediktorer (p-värde < 0.05):
> print(significant_predictors)
 [1] "(Intercept)"               "Ålder"                     "Miltal"                    "Hästkrafter"
 [5] "Bränslediesel"             "Bränsleel"                 "Bränslemiljöbränsle/hybrid" "SäljarePrivat"
 [9] "BiltypFamiljebuss"         "Växellådamanuell"          "ModellgruppKompaktbilar"   "ModellgruppLyxmodeller"
[13] "ModellgruppMinibussar"     "ModellgruppÖvriga"         "ModellgruppSUV:ar"
```

*Bildtext: Utskrift som visar prediktorer i Model 2 med p-värden under 0.05 och bekräftar vilka bilattribut som signifikant påverkar log(försäljningspris). Källa: Egen analys i R.*

Figur A19: Konfidensintervall för Model 2:s prediktorer

```
Konfidensintervall (95%) för koefficienterna:
> confint_final <- confint(final_model, level = 0.95)
> print(confint_final)
                                  2.5 %        97.5 %
(Intercept)                1.231897e+01  1.255675e+01
Ålder                     -9.798935e-02 -8.789815e-02
Miltal                    -3.062829e-05 -2.406797e-05
Hästkrafter                3.746399e-03  4.503732e-03
Bränslediesel              8.945135e-02  1.687245e-01
Bränsleel                 -3.557302e-01 -2.200864e-01
Bränslemiljöbränsle/hybrid -1.437521e-01 -4.047602e-02
SäljarePrivat             -1.629332e-01 -8.260254e-02
BiltypFamiljebuss          1.393067e-02  2.437559e-01
BiltypHalvkombi           -1.424878e-01  4.049758e-02
BiltypKombi               -1.709487e-01  6.125382e-05
BiltypSUV                 -2.070601e-01  1.374545e-03
Växellådamanuell          -1.711814e-01 -8.977631e-02
ModellgruppKompaktbilar    6.899964e-03  1.081612e-01
ModellgruppLyxmodeller     4.925653e-02  3.041183e-01
ModellgruppMinibussar      1.488317e-01  3.812918e-01
ModellgruppÖvriga          8.927996e-02  4.041022e-01
ModellgruppSUV:ar          9.809124e-02  2.390663e-01
```

*Bildtext: Utskrift som visar 95 % konfidensintervall för koefficienterna i Model 2. Källa: Egen analys i R.*

Figur A20: Konfidens- och prediktionsintervall

```
Konfidensintervall (95%) för medelvärdet av predikterade priser i SEK:
> print(confidence_intervals_sek)
       fit       lwr       upr
1 166697.5 156364.26 177713.56
2 155037.1 143108.00 167960.54
3  66749.4  63608.53  70045.37
4 243226.7 229084.00 258242.61
5 251347.9 238349.09 265055.57
> cat("\nPrediktionsintervall (95%) för individuella predikterade priser i SEK:\n")

Prediktionsintervall (95%) för individuella predikterade priser i SEK:
> print(prediction_intervals_sek)
       fit       lwr      upr
1 166697.5 107020.64 259651.3
2 155037.1  99275.69 242118.7
3  66749.4  42939.42 103762.1
4 243226.7 156242.12 378638.3
5 251347.9 161599.35 390940.6
```

*Bildtext: Utskrift som visar 95 % konfidensintervall (medelpris) och prediktionsintervall (individuellt pris) i SEK för de första fem observationerna i testdatan. Källa: Egen analys i R.*

Tabell A1: Jämförelse av prestandamått för Model 1, Model 2 och Model 3

| **Modell** | **CV10 Mean RMSE (log-skala)** | **CV10 Mean R^2** | **SD RMSE** | **SD R^2** |
|---|---|---|---|---|
| Model 1 | 0.2491 | 0.9022 | 0.0565 | 0.0401 |
| Model 2 | 0.2478 | 0.9031 | 0.0564 | 0.0394 |
| Model 3 | 0.2489 | 0.9032 | 0.0529 | 0.0364 |

*Bildtext: Tabell som jämför prestandamått (t.ex. CV10 $R^2$, RMSE i log-skala) för Model 1, Model 2, och Model 3 på träningssetet. Källa: Egen analys i R.*

Tabell A2: Medelålder och modellgrupp per bränsletyp

| **Bränsle** | **Medelålder (år)** | **Kompaktbilar (%)** | **SUV:ar (%)** |
|---|---|---|---|
| Bensin | 10.3 | 69.3 | 17.2 |
| Diesel | 8.91 | 23.4 | 25.4 |
| El | 3.77 | 43.5 | 56.5 |
| Miljöbränsle/hybrid | 8.00 | 32.2 | 11.2 |

*Bildtext: Tabell som visar medelålder och fördelning av modellgrupper (Kompaktbilar och SUV:ar) per bränsletyp, baserat på rådatan. Elbilar är betydligt nyare och oftare SUV:ar, vilket bidrar till deras högre medelpris i rådatan (se avsnitt 4.3). Källa: Egen analys i R.*

# Källförteckning


- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning: With applications in R (2nd ed.). Springer. (Corrected printing, 2023).

- SCB (2025). Nyregistrerade personbilar efter län och kommun samt drivmedel. Månad 2006M01 - 2025M03. Statistiska Centralbyrån. Hämtad från https://www.statistikdatabasen.scb.se/pxweb/sv/ssd/START__TK__TK1001__TK1001A/PersBilarDrivMedel/