När gruppen är klar med datainsamlingen så skall du besvara följande frågor kortfattat (ha ett kapitel i rapportens metod del som t.ex. heter ”Datainsamling”):

1. Vem du har arbetat i grupp med?

Filip, David, Richard och Erik.

2. Hur har ni i gruppen arbetat tillsammans?

Vi tog först beslut om vilken data som skulle samlas in sen hade vi avstämningar de första dagarna via teams-chat. Arbetet fördelades så att alla gjorde lika mycket och bidrog med den kunskapen, erfarenhet man kunde inom ämnen som bilar, verktyg, formattering etc.

3. Vad var bra i grupparbetet och vad kan utvecklas?

Tyckte det fungerade bra, alla tog ansvar och drog sitt strå till stacken. Hade tur med gruppen, och alla ville lyckas väl vilket speglar engagemang. Ger stöd när det behövs.

4. Vad är dina styrkor och utvecklingsmöjligheter när du arbetar i grupp?

Jag är van att båda va i grupp och leda grupp så försöker vara lyhörd, visa respekt, våga ta plats, utan att köra över någon.  
Inom just detta ämnet så är jag ju inte alls erfarenhet, dvs R och kanske kan lita på min intuition mer.

5. Finns det något du hade gjort annorlunda? Vad i sådana fall?

Vi hade inte samlat in data med bestämt pris-spann, utan låtit den variabeln obestämd, samt kanske mer om utrustning, plats etc.

Gruppmässigt så hade jag inte gjort något annorlunda, tyckte den fungerade bra och alla stöttar och hjälper till.

(1) Teoretiska frågor

Besvara följande teoretiska 7 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.  
Man plottar sina sample quanteties emot dess värde från standard normal kurvan och ska få en linje då det ska vara ett högt linjert samband mellan dessa, en pretty darn straight line, helt enkelt. Vill man va säker kan man testa med Shapiro-Wilk test med.

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?

Hej Karin, I maskininlärning ligger fokus på att skapa modeller som är bra på att förutsäga nya data, ibland och man behöver inte förstå varför modellen fungerar som den gör, hur relationen mellan de olika parametrarna hänger ihop.

I statistisk regressionsanalys syftar man inte bara till att förutsäga nya data utan också att förstå relationerna mellan variabler, och på så sätt kan använda modellen på ett bra sätt när några parametrar ändras, som ex variabler som har med omvärlden att göra och man behöver förstå mer teoretiskt när det är läge ex att köpa aktier, eller när det är läge att sälja sin bostad bil etc.

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

**Konfidensintervall** används för att uppskatta osäkerheten kring ett parameterestimat i en statistisk modell. Till exempel, om man har en linjär regressionsmodell där man estimerar sambandet mellan två variabler, så ger ett 95% konfidensintervall för regressionskoefficienten det intervallet där vi är 95% säker på att den sanna koefficienten ligger. Detta handlar alltså om osäkerheten kring själva modellparametern.

**Prediktionsintervall** å andra sidan ger en uppskattning av var nya data (inte de data som använts för att träna modellen) förväntas ligga på med en viss sannolikhet. Det tar inte bara hänsyn till osäkerheten i parameterestimaten, utan också den inneboende variabiliteten i data. Till exempel, om man använder samma regressionsmodell för att förutsäga ett värde för en ny observation, så ger prediktionsintervallet det område där denna nya observation förväntas falla med en viss sannolikhet, exempelvis 95%.

4. Den multipla linjära regressionsmodellen kan skrivas som:

Y = β0 + β1x1 + β1x2+ . . . + βpxp + ε .

Beta (β) parametrarna i en regressionsmodell tolkas som koefficienter som anger den förväntade förändringen i den beroende variabeln (Y) för varje enhets-förändring i den oberoende variabeln (x), med hänsyn till andra variabler i modellen som inte förändras

5. Din kollega Hassan 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?

Hej Hassan. Kul att du frågade!   
Logiken här är att BIC straffar modeller som har för många parametrar (överanpassning). Här kan man jämföra modellers BIC och gå efter det som är lägst.Man behöver ej validera data, utan kan gå vidare med den modell som har bäst för att prediktera.

6. Förklara algoritmen nedan för ”Best subset selection”

Best Subset Selection är en teknik i regressionsanalys där man väljer den bästa kombinationen av förklarande variabler (features) som ger den mest prediktiva modellen. Algoritmen fungerar i grova drag så här:

Start: Börja med att betrakta alla möjliga kombinationer av variabler.

Beräkning: För varje möjlig kombination av variabler, tränar man en modell och beräknar mått som BIC, AIC eller justerad R-kvadrat.

Optimering: Välj den modell som har bäst balans mellan god anpassning (hög R-kvadrat) och modellens enkelhet (färre variabler, låga värden på BIC/AIC).

7. Ett citat från statistikern George Box är: “All models are wrong, some are useful.”

Förklara vad som menas med det citatet.

Att modellerna man skapar sällan är perfekta, men det betyder inte att de inte är användbara. När vi skapar modeller så handlar det inte om att leva efter dem men förstå när de kan hjälpa oss på vägen, och vilken typ av fråga man kan få svar på från dem