Frågor om grupparbete:

1. Vem du har arbetat i grupp med?
   Keiket, Jakob, Robert och Melissa
2. Hur har ni i gruppen arbetat tillsammans?
   Vi träffades först för att diskutera vilken data och vilka parametrar vi ville samla in, sedan samlade vi in ett begränsat antal observationer och träffades dagen därpå för att utföra en POC, vi alla tyckte att datan såg bra ut och vi bestämde oss därför för att samla in mer data av samma typ.
3. Vad var bra i grupparbetet och vad kan utvecklas?
   Vi hade väldigt bra diskussioner där alla kunde få säga sin åsikt och vi kunde diskutera fram och tillbaka, vi hade kunnat ha det lite mer strukturerat för att snabbare kunna lösa problem och se till att alla får fram sina åsikter.
4. Vad är dina styrkor och utvecklingsmöjligheter när du arbetar i grupp?
   Jag anser att jag kan bidra med ett unikt perspektiv men skulle kunna förbättras på att se det från andras perspektiv och formulera mina tankar och förslag bättre.
5. Finns det något du hade gjort annorlunda? Vad i sådana fall?
   Jag tyckte att grupparbetet gick väldigt bra och är nöjd. Det enda som hade kunnat förbättras hade varit lite mer struktur under diskussionerna då jag tror detta hade förbättrat och snabbat på processen.

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.
   Syftet med en Quantile-Quantile plot(QQ-plot) är att tolka om datan är normalfördelad. Detta görs genom att jämföra kvantilerna för datan och de teoretiska kvantiler från en normalfördelning med samma storlek som datan. Om värdena är nära varandra resulterar detta i en rak linje vilket tyder på att datan är normalfördelad.
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?
   Inom maskinlärning så ämnar man främst att skapa modeller som utifrån rätt data kan uppnå så precisa prediktioner som möjligt, även fast man inom statistik regressionsanalys också vill uppnå så bra resultat som möjligt läggs det även fokus på att förstå modellen och variabler, man vill få en djupare förståelse för vad som påverkar prediktionerna och varför.
3. Vad är skillnaden på ”konfidensintervall” och ”prediktionsintervall” för predikterade värden?
   Prediktionsintervall används för att prediktera utfallet av en ny observation medans Konfidensintervall baseras på snittet. Prediktionsintervall inkluderar epsilon i beräkningen vilket leder till att prediktionsintervall alltid är bredare.
4. Den multipla linjära regressionsmodellen kan skrivas som: $Y=\beta_0+\beta_1x_1+\beta_1x_2+\ldots+\beta pxp+\varepsilon$. Hur tolkas beta parametrarna?
   Vad är effekten på Y när en beta parameter ändras givet att de andra beta parametrarna är fixa.

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?

Det stämmer att man utifrån samma data med hjälp av till exempel BIC kan skatta testfelet. Logiken är att man skattar ett testfel utifrån träningsdatan som straffar komplexitet vilket förhindrar overfitting.

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 provar en modell för alla möjliga kombinationer av variabler väljer den bästa modellen baserat på måttet RSS.

7. Ett citat från statistikern George Box är: *“All models are wrong, some are useful.”*Förklara vad som menas med det citatet.
Att ingen model kommer kunna vara 100% korrekt, alla modeller bygger på antaganden. Trots detta är många modeller användbara eftersom de kan ofta ge oss insikt i datan och göra prediktioner för framtiden som är användbara.

Självutvärdering

1. Utmaningar du haft under arbetet samt hur du hanterat dem.
   Det har uppstått en del utmaningar under arbetet, främst med variabel selektion och dummy variables men även med modell utvärdering och förbättring. Jag hantera problem genom att kolla olika forum och källor online där olika metoder har provats för att lösa olika problem. Jag har sedan provat de olika metoderna och använt de bästa i mitt slutgiltiga projekt. Mycket har jag även löst genom att tänka logiskt och försöka motivera det som till exmpel konvertering av motorstorlek till liter från kubikcentimeter.

2. Vilket betyg du anser att du skall ha och varför.
   Jag anser att jag ska ha VG eftersom jag har löst flera problem med metoder från regressionsanalys och jag har gjort detta med hög säkerhet. Jag har även i min rapport lyft potentiella problem och förhinder som har uppstått och hur de har hanterats. Utöver detta har jag använt mig av SCB API för att integrera data i min rapport som var ett del av VG kraven.

3. Något du vill lyfta fram till Antonio?
   Att det var en utmanande uppgift som krävde mycket kritiskt tänkande.