**Kunskapskontroll R Programmering – Del 1 Datainsamling**

**Uppgift**Samla in data från Blocket(https://www.blocket.se) om bilar och lagra den i Excel. Datan ska sedan användas för att utföra en regressionsanalys. 

**Ni kommer göra en modell, vad är syftet med modellen och vilken data behövs för det?**

Syfte: Prediktera pris på elbilar SUV (Sport Utility Vehicle) som ska säljas i Skåne – automatisk prissättning via regressionsmodellering.   
Typ av data: De olika bilarnas utvalda features. 

**Vilken typ av fordon vill vi modellera? *Exempelvis kan det vara problematiskt om hälften är exklusiva bilar såsom Ferrari och andra hälften vanliga bilar såsom Mazda.*** Fordon att modellera: Elbilar som lagts in och kategoriserats på Blocket som SUV.

Elbilar som rullat minst 50 mil då vi valt att inte inkludera ny producerade bilar som ej tagits i trafik.

Begränsningen är gjord till el bilar i kategori SUV, som säljs av bilförmedlare i Skåne, leasingbilar exkluderas.

**Säkerställ att datan ni samlar in går att läsa in i R och att det blir som ni tänker er. Gör alltså en ”Proof of Concept” (POC).**    
**Göra en ”testkörning” när alla samlat in vars ca 20 bilar?**

Delar av datan är testkörd och fungerar både med Exel och CSV.

**Planering och förberedelse**

**Förstå uppgiften och kraven för datainsamlingen**

Vi tror oss förstått uppgiften för datainsamlingen och att kraven är acceptabla.

Tillägas skall att vi har resonerat kring vår områdes begränsning som endast innefattar Skåne.

**Säkerställa att vi har tillgång till de verktyg som behövs, som exempelvis Excel.**

Alla i gruppen har tillgång, samt erforderlig kunskap i Excel för uppgiften.

**Skapa en gemensam plan i grupp om hur vi delar upp arbetet och kommunicerar under processen.**

Insamling av datan har fördelats jämlikt på alla utifrån miltal, där vi har tagit fasta på stabiliteten i miltalet på utannonserades objekt.  
Insamlingen av datan har vi gjort i separata filer som sedan har sammansatts i en fil.  
Vi kommunicerar löpande via Teams.

|  |  |
| --- | --- |
| **Miltal** | **Antal observationer** |
| 50 - 800 | 97 |
| 800-1700 | 102 |
| 1701 - 3000 | 93 |
| 3001-5500 | 90 |
| < 5501 | 85 |

**Skapa en Excel-fil med kolumnrubriker för varje attribut som ska samlas in samt fyll i ett exempel så att alla i gruppen fyller i på samma sätt.**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Annons-URL** | **Data inhämtad (datum)** | **Märke** | **Modell** | **Modellår** | **Miltal** | **Hästkrafter** | **Färg** | **Pris** |
| https://www.blocket.se/annons/1000556260 | 11/04/2024 | Jeep | Avenger | 2023 | 50 | 157 | Gul | 449 900 |

**Datainsamling**

**Vi har avgränsat annonssökningen på Blocket.se enligt nedan val i kronologisk ordning, enligt följande:**    
1. Katogori: fordon, bilar 133 237 st   
2. Drivmedel: El 16 876 st   
3. Biltyp: SUV 7 634 st   
4. Miltal: < 50 mil 4 144 st   
5. Ägandeform: köpa 4 081 st   
6. Säljare, Företag 3 822 st   
7 .Område, Skåne 483 st

*Data inhämtad från Blocket.se 2024-04-11 k. 13.30*

 *2024-04-11 k. 13.30*

[https://www.blocket.se/bilar/sok?filter=%7B%22key%22%3A%22car.details.mainPropellant.fuel%22%2C%22values%22%3A%5B%22El%22%5D%2C%22type%22%3A%22string%22%7D&filter=%7B%22key%22%3A%](https://www.blocket.se/bilar/sok?filter=%7B%22key%22%3A%22car.details.mainPropellant.fuel%22%2C%22values%22%3A%5B%22El%22%5D%2C%22type%22%3A%22string%22%7D&filter=%7B%22key%22%3A%25)22car.details.chassi%22%2C%22values%22%3A%5B%22SUV%22%5D%2C%22type%22%3A%22string%22%7D&filter=%7B%22key%22%3A%22car.details.milage%22%2C%22range%22%3A%7B%22start%22%3A%2250%22%2C%22end%22%3A%22%22%7D%2C%22type%22%3A%22int%22%7D&filter=%7B%22key%22%3A%22deal.ownershipType%22%2C%22values%22%3A%5B%22K%C3%B6pa%22%5D%2C%22type%22%3A%22string%22%7D&filter=%7B%22key%22%3A%22deal.location.regionName%22%2C%22values%22%3A%5B%22Sk%C3%A5ne%22%5D%2C%22type%22%3A%22nested%22%7D&filter=%7B%22key%22%3A%22deal.owner.type%22%2C%22values%22%3A%5B%22F%C3%B6retag%22%5D%2C%22type%22%3A%22string%22%7D

**Samla in data från annonserna som uppfyller kraven -> märke, modell, modellår, pris, miltal, hästkrafter och färg.**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Annons-URL** | **Data inhämtad (datum)** | **Märke** | **Modell** | **Modellår** | **Miltal** | **Hästkrafter** | **Färg** | **Pris** |

**3. Organisera data i Excel**  
**Fyll i den insamlade datan i Excel-filen enligt de specificerade kolumnrubrikerna.**

Data är insamlad enligt ovan modell under perioden 11-14 april 2024 från Blocket.se

**4. Kontrollera datan**  
**Granska den insamlade datan för att säkerställa att den motsvarar förväntningarna och att inga felaktiga värden finns.**

**Kontrollera att alla kolumner är fyllda med relevant information.**

Datan är sammanställd, kontrollerad och rensad.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**5. Spara datan och förbered för analys i R**  
**Spara Excel-filen som en CSV-fil för inläsning i R. (För mer flexibilitet?)**

**6. Läsa in datan i R och utforska den**  
**Läs in csv-filen i R.   
Utforska datan genom tex head(), tail() och summary() för att kontrollera datakvaliteten och få en överblick över datan.**

**7. Utvärdering**  
**Eventuella problem och lösningar som uppstod?    
Utvärdera om det blev framgångsrikt och om datan är lämplig för att utföra regressionsanalysen.   
Identifiera eventuella förbättringsområden inför den fullständiga implementationen av projektet.**

***- Vilken typ av data skall vi samla in?***

***- Kan man göra några kontroller så datan är ”rimlig”?***    
 Priskontroll?   
 Miltalskontroll?

***Övrigt:***   
 Tankar om miljöpåverkan. Möjliga variabler som påverkar: pris på el/bensin/diesel men också laddningsinfrastruktur. Miljöstöd?   
   
 Längre fram i projektet kan extern data nyttjas för att visa andra variabler som påverkar men också för att skapa en övertygande argumentation.

Genom extern data kan vi även motivera varför arbetet är intressant och dessutom analysera värdet av att skapa modeller för prissättning.

**Projektaktiviteter**   
Onsdag under lektion – Gruppindelning och första möte. Diskussion om mål och planering av vårt arbete framåt.   
Torsdag kl 9:30 - Gruppmöte. Avstämning. Nytt möte bestämdes så att alla skulle få tid att reflektera över vad vi kommit fram till och hinna arbeta framåt för att upptäcka eventuella frågor eller bekymmer eller för den delen framsteg.   
Torsdag kl 14:30 – Gruppmöte. Kontroll att vi är överens och redo för att påbörja insamling av data. Målsättning att vara klara med datainsamling senast måndag 2024-04-15.   
Möte kl 14:30 måndag 2024-04-15. Kontroll av insamlad data.