Teoretiska Frågor

1. Beskriv kort hur en relationsdatabas fungerar.

Relationsdatabaser består av tabeller som består av kolumner(olika egenskaper av data) och rader(olika objekt), och olika tabeller kopplar till varandra genom "primery key" (unika identifierare) och "foreign keys"(unika identifierare i en annan tabel). Relationsdatabas är ett effektivt sätt att hantera och hämta stora mängder och komplicerad data.

2. Vad menas med ”CRUD” flödet?

Det är Create(att skapa/lägga till nya data till databasen), Read(att hämta och läsa data från databasen). Update(att uppdatera/ändra vissa data i databasen), och Delete(att ta bort data från databasen)

3. Beskriv kort vad en ”left join” och ”inner join” är. Varför använder man det?

Left join: ett sätt att kombinera tabeller. "Left join" gör att man behåller all data från vänster tabell och försöker matcha med rows från höger tabell, där man inte hittar matchande data i höger tabell står det NULL. inner join: ett annat sätt att kombinera tabeller. Det visar bara data som kan hittas i båda tabeller.

4. Beskriv kort vad indexering i SQL innebär.

Indexering är ett sätt att kunna söka data i en databas snabbare. Det skapar en struktur så att man inte behöver söka igenom hela databasen.

5. Beskriv kort vad en vy i SQL är.

Vy lagrar inte data, istället det kan betraktas som virtuella tabeller som sparar "queries" så att när man vill nästa gång snabbt ta fram data med samma query, kan man genomföra det via att klicka på vy.

6. Beskriv kort vad en lagrad procedur i SQL är

Som vy, kan en lagrad procedur processera databaser mer effektivt, men den kan spara ner komplexa operationer såsom datainmätning, datamodifikation, och andra mer komplexa SQL koder.

Rapport

1. AdventureWorks2022

AdventureWorks2022 är en omfattande databas som består av fyra huvudschema: Person, Production, Purchasing och Sales. Dessa schema representerar olika affärsområden såsom personalresurser, produktion, inköp och försäljning. Databasen innehåller totalt 72 tabeller fördelade på dessa fyra kategorier.

Inom schema för Produktion finns det 25 tabeller som innehåller detaljerad information om produkter, såsom produktnamn och ID, prisutveckling, lagerstatus, kostnad mm.

Person schema omfattar 14 tabeller som förser information om anställdas personuppgifter, anställningsavdelning, geografiska regioner mm.

Purchasing schema består av 5 tabeller med data relaterad till inköp, logistik, leverantörsinformation, inköpshistorik mm.

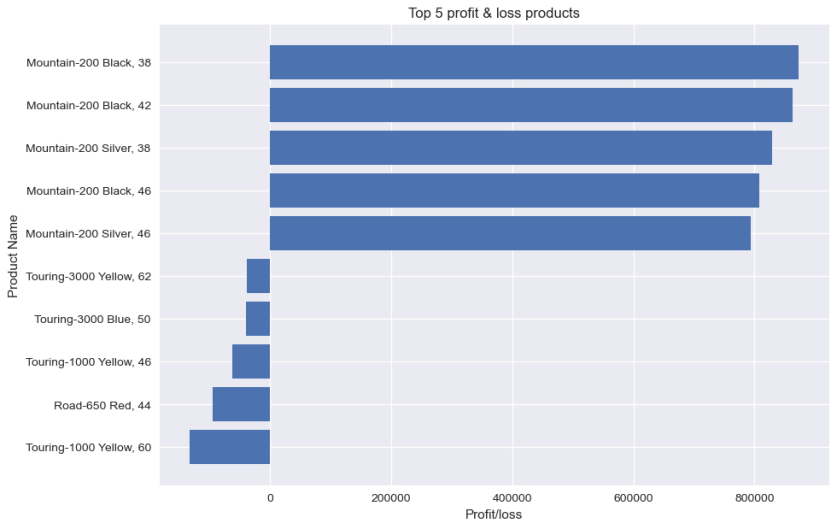
Inom försäljningsschema finns det 19 tabeller som ger en överblick över viktiga affärsaspekter, inklusive orderdatum, prissättning, vilka produkter och i vilken mängd som beställts, försäljningsregioner och valutor, kundinformation, osv.

2. Djupdykning i försäljningsdata & statistisk analys

Fråga 1: Vilka produkter har genererat högst respektive lägst vinst?

För att besvara denna fråga, har jag skapat en vy som visar varje OrderLine med respektive produkter, pris, kostnad (som motsvarar respektive försäljningsdatum), vinst och marginal.

Därefter utförde jag en query som summerar den totala vinsten för varje produkt, vilket resulterade i en lista över de fem mest och minst lönsamma produkterna. Se bild nedan (Top 5 profit & loss products).



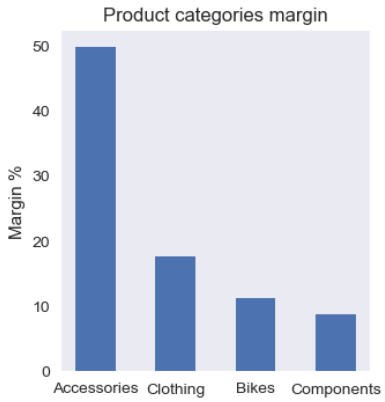
Diagrammet visar att ”Mountain Bike 200” (i olika storlekar) är de produkterna som genererat mest vinst för företaget under perioden 2011-2014. Å andra sidan finns det en rad produkter som gör negativ vinst, dvs förlust, där de flesta är touring bikes. En närmare granskning av databasen (se bild nedan) avslöjar att alla de fem produkterna med förlust deltog i någon form av utförsäljning eller lagerrensning, ofta p.g.a. överlager eller införande av en ny produkt, vilket resulterade i avsevärda rabatter (mellan 15% och 30%). Detta förklarar den negativa vinsten för dessa produkter.

	ProductID	Description	DiscountPct	Type
0	762	No Discount	0.00	No Discount
1	762	Volume Discount 11 to 14	0.02	Volume Discount
2	762	Volume Discount 15 to 24	0.05	Volume Discount
3	762	Road-650 Overstock	0.30	Excess Inventory
4	954	No Discount	0.00	No Discount
5	954	Volume Discount 11 to 14	0.02	Volume Discount
6	954	Volume Discount 25 to 40	0.10	Volume Discount
7	954	Touring-1000 Promotion	0.20	New Product
8	957	No Discount	0.00	No Discount
9	957	Volume Discount 11 to 14	0.02	Volume Discount
10	957	Volume Discount 15 to 24	0.05	Volume Discount
11	957	Touring-1000 Promotion	0.20	New Product
12	965	No Discount	0.00	No Discount
13	965	Volume Discount 11 to 14	0.02	Volume Discount
14	965	Touring-3000 Promotion	0.15	New Product
15	979	No Discount	0.00	No Discount
16	979	Volume Discount 11 to 14	0.02	Volume Discount
17	979	Volume Discount 15 to 24	0.05	Volume Discount
18	979	Touring-3000 Promotion	0.15	New Product

Observationer visar också att alla produkter med högst och lägst vinst är cyklar, vilket väcker frågor om cyklars övergripande vinstmarginaler och hur dessa står sig i jämförelse med andra produktkategorier, vilket leder oss till fråga 2.

Fråga 2. Vad är vinsten och marginalerna för de olika produktkategorierna?

För att besvara denna fråga har en query genomförts som beräknar marginal och vinst för varje produktkategori. bilder nedan (Product categories margin & Product categories profit).

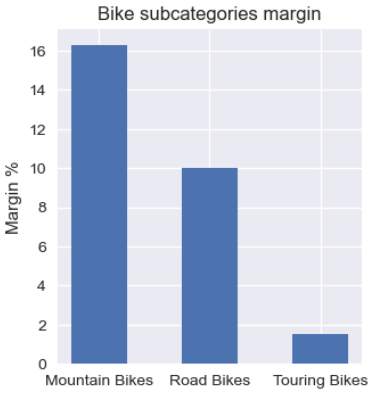


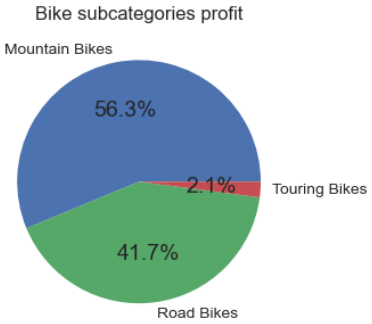
Diagrammet på vänster visar att marginalen för cyklar ligger omkring 10%, vilket är en relativt låg marginal jämfört med andra kategorier såsom tillbehör (Accessories) och kläder (Clothing). Trots detta står cyklar för 83,8% av företagets totala vinst, vilket tyder på att en betydande del av försäljningen utgörs av cyklar. För att maximera vinsten kan det vara lönsamt att i framtiden utforska mer på försäljning av tillbehör.

För att få en djupare insikt i vilka typer av cyklar som säljs och hur dessa subkategorier presterar med avseende på marginal och vinst, övergår vi nu till nästa frågeställning.

Fråga 3. Hur ser marginalen och vinsten ut för olika subkategorier av cyklar?

För att besvara denna fråga har jag genomfört en förfrågan (query) som beräknar



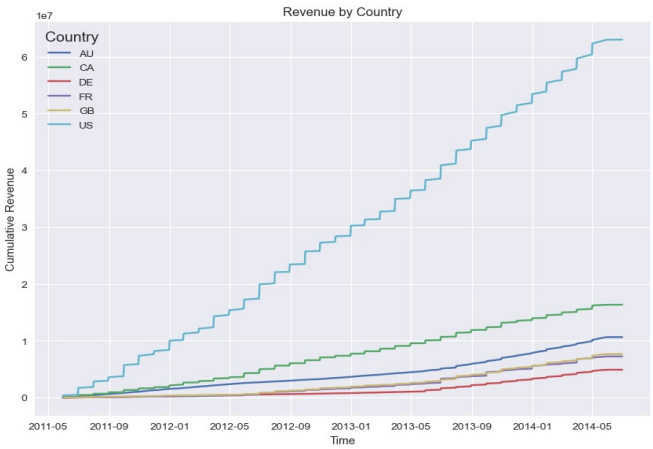


marginalen och vinsten för varje cykelsubkategori. Se bild nedan (Bike subcategories margin & Bike subcategories profit).

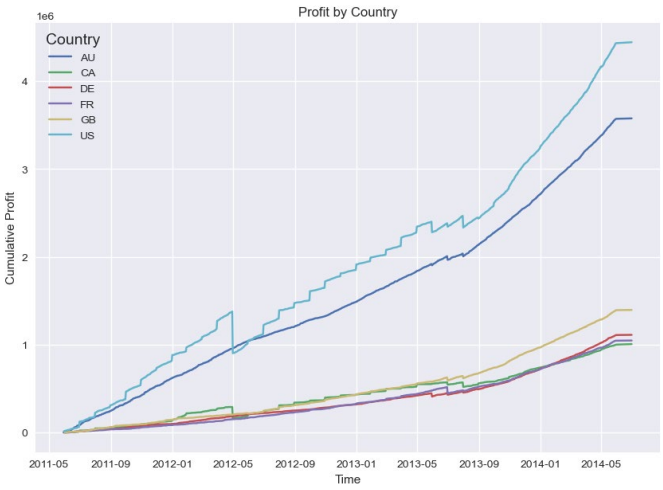
Diagrammen visar att mountain bikes har den högsta marginalen medan touring bikes har den lägsta. Detta bekräftar tidigare resultat från fråga 1 där mountainbikes utgjorde de fem mest lönsamma produkterna och touring bikes de flesta av de minst lönsamma. Pie chart på höger visar att mountain bikes och Road Bikes står för majoriteten av vinsten inom cykelförsäljningen. Det kan vara värt att undersöka om det finns behov av att justera lager- och försäljningsstrategierna för touring bikes för att öka deras lönsamhet, eftersom de nuvarande låga vinstsiffrorna delvis beror på lagerrensningar och övergång till nya produkter.

Fråga 4. Hur ser utvecklingen av intäkter och vinster ut för olika regioner?

För att undersöka trenden för intäkter och vinster i olika regioner, har en SQL-vy förberetts där daglig vinst och intäkt listas för varje region. För att skapa grafer som visar dessa trender, har en förfrågan (query) använts för att beräkna kumulativ vinst och intäkt dag för dag. Sedan har ett diagram genererats för att illustrera intäktsutvecklingen i olika regioner. Se bild nedan.



Diagrammen visar en utveckling över åren i alla regioner, där den amerikanska marknaden (US) uppvisar en snabbare tillväxt jämfört med andra regioner. Kanada och Australien ligger bakom USA, och de tre europeiska länderna uppvisar de lägsta vinsterna.



En ytterligare graf har tagits fram för att jämföra vinstutvecklingen i olika regioner. En intressant observation från graferna är att Kanada, trots att det ligger på andra plats när det gäller intäkter, har en relativt låg vinst, medan Australien, som har lägre intäkter jämfört med Kanada, ändå har mycket högre vinst.

Innan vi drar några slutsatser, behöver vi genomföra en t-test för att bekräfta att det finns en signifikant skillnad i intäkter och vinster mellan Kanada och Australien. Beräkningen kommer att baseras på genomsnittlig daglig intäkt och vinst i de två länderna. P-värdet för skillnaden mellan genomsnittlig daglig intäkt för de två länderna är 0.028, vilket är lägre än 0.05, och det tyder på en signifikant skillnad i genomsnittlig daglig intäkt. Därefter jämförs vinster. P-värdet är extremt lågt, nära 0, vilket indikerar en signifikant skillnad i genomsnittlig daglig vinst mellan länderna.

Denna slutsats kan stärka vidare analys och jämförelser inom företagsstyrning och försäljningsstrategier mellan de två länderna för att identifiera förbättringsområden. Det kan också bidra till att hitta globala strategier som kan maximera vinsten i alla regioner.

Executive Summary

I denna analys av AdventureWorks2022 har vi utforskat vinstgenerering av olika produkter, produktkategorier och regioner. Analysen avslöjar att "Mountain Bike 200" är den mest lönsamma produkten medan touring bikes är minst lönsamma, ofta på grund av utförsäljningar. Cyklar, trots en lägre marginal jämfört med andra kategorier som tillbehör och kläder, står för majoriteten av företagets vinst. Mountain bikes har högst marginal bland cykelsubkategorierna.

En regional analys visade att USA-marknaden hade en snabbare vinsttillväxt jämfört med andra regioner. Trots högre intäkter i Kanada än i Australien var vinsten lägre. Ett t-test bekräftade signifikanta skillnader i intäkter och vinster mellan Kanada och Australien, vilket vid vidare analys kan tyda på viktiga områden för strategiska förbättringar.

Kommentar

Det har uppstått flera stora och mindre utmaningar under utförandet av denna uppgift, vilka har bidragit till att fördjupa min förståelse för Python och SQL. Nedan följer några exempel:

1. När frågeställning blir för komplex är det bättre att först skaffa en vy, sedan en ytterligare query
2. Lärde mig följande kodning:
    a. Produktkostnader varierar över tid, behöver matcha med rätt pris enligt OrderDate i query
    b. Skapa en pivot dataframe för att underlätta generering av grafer
    c. använder for loop för att skapa flera linjer i samma diagram
    d. hantera situationer där vissa datum saknar data för vissa regioner, vilket leder till diskontinuerliga linjer. Lösning är att fylla på/ersätta NaN i df.
3. Vilka slutsatser man (inte) kan dra från olika resultat.

För att tackla dessa utmaningar har jag återvänt till kursmaterial från både denna och föregående kurs, diskuterat med vännen som arbetar som business analyst, sökt information på Google och frågat ChatGPT. Sammanfattningsvis har processen varit mycket lärorik.

Jag kan nog få VG i kursen, för jag byggde några mer utmanande vyer och queries jämfört med de vi gick igenom i kursen och har skaffat en fullständig rapport som visade bra förståelse av databasen.

Jag önskar att jag hade börjat med liknande, mindre projekt tidigare i kursen för att praktisera Python mer, eftersom jag kände att jag hade glömt mycket om Python när jag påbörjade uppgiften.