Kunskapskontroll SQL

Teoretiska frågor

1. Beskriv kort hur en relationsdatabas fungerar.

2. Vad menas med “CRUD”-flödet?

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

4. Beskriv kort vad indexering i SQL innebär.

5. Beskriv kort vad en vy i SQL är.

6. Beskriv kort vad en lagrad procedur i SQL är.

Teoretiska frågor svar:

1. Beskriv kort hur en relationsdatabas fungerar.

En relationsdatabas är en typ av databas som strukturerar och lagrar information i tabeller bestående av rader och kolumner. Datan har ofta en relation mellan varandra. Varje tabell fokuserar på en specifik typ av data, till exempel kunder, produkter eller beställningar, och kopplingarna mellan dessa data hanteras med hjälp av primärnycklar och främmande nycklar.

2. Vad menas med “CRUD”-flödet?

CRUD står för Create, Read, Update och Delete. Dessa representerar de fyra huvudsakliga funktionerna för att arbeta med data i en databas eller ett informationssystem. Dessa fyra funktioner används för att skapa, läsa, ändra och radera data och används för att interagera med och hantera information på ett systematiskt sätt.

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

En **left join** och en **inner join** används för att lägga samman data från två eller fler tabeller i en databas med hjälp av ett gemensamt fält.

- **Left join**: Används för att hämta alla rader från den vänstra(första) tabellen och matchande rader från den högra tabellen. Om inga matchade värden återfinns retureras NULL värden från den högra tabellen. Detta används när man vill få fram alla värden även fast det inte finns några matchande värden i den andra tabellen.

- **Inner join**: Används när man vill hämta rader då det finns en mathcning mellan två tabeller. Rader utan matchningar exkluderas. Detta är användbart när man enbart vill ha data som existerar i båda tabellerna.

4. Beskriv kort vad indexering i SQL innebär.

Indexering används så att man kan hämta data ur databasen på ett snabbare och mer effektivt sätt. När en tabell indexeras, skapar databasen en separate struktur som lagrar de indexerade kolumnerna tillsammans med respective rader. När man sedan gör queries, kan databasen snabbt hämta datan med hjälp av indexeringen och hämta de relevanda raderna istället för att söka igenom hela tabellen.

5. Beskriv kort vad en vy i SQL är.

En vy fungerar som en virtuell tabell baserat på en query. Den sparar ingen data utan visar data som är sparad i andra tabeller. Vyer används ofta för att förenkla komplexa frågor genom att definera dem till ett enda objekt, förbättra läsbarheten eller begränsa åtkomsten till viss data genom att visa enbart relevanta kolumner eller rader.

6. Beskriv kort vad en lagrad procedur i SQL är.

Lagrade procedurer i SQL är redan förskriva funktioner som är sparade i databasen. Dessa kan användas för att utföra specifika uppgifter såsom att skapa queries, updatera data eller liknande. Eftersom dessa “funktioner” redan är sparade i databasen, gör det att processen går fortare att utföra.

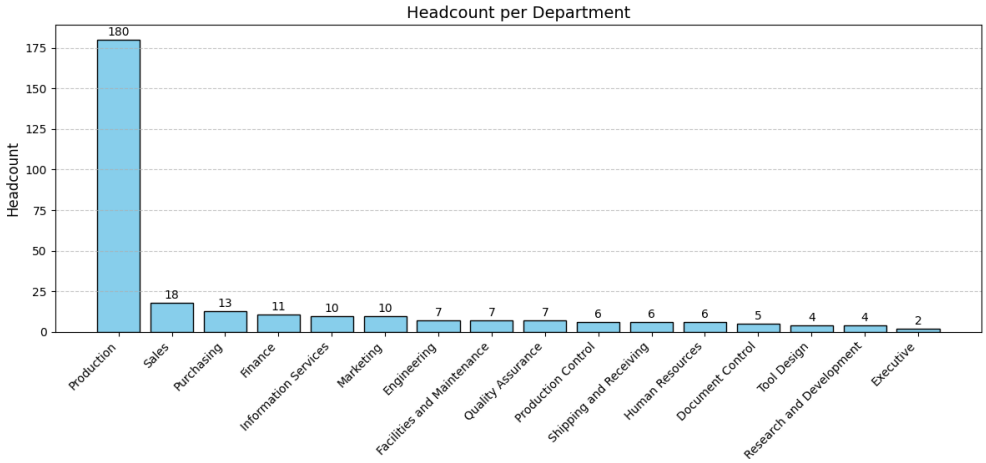
Programmeringsuppgift

Företaget AdventureWorks är ett företag som säljer cyklar, cykeltillbehör, cykelkomponenter och cykelkläder. Databasen består av fem olika scheman, där varje schema innehåller egna tabeller för respektive schema. Dessa scheman är: HumanResources, Person, Production, Purchasing och Sales.

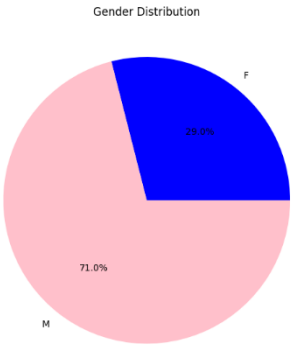
Vi kan börja med att analysera alla fem scheman för att beskriva datan som finns.

HUMAN RESOURCES

Från **HumanResources** schemat kan vi utläsa att företaget har totalt 296 anställda fördelade över 16 kategorier. Flest antal anställda (180st) har avdelningen ”Production”. Minst antal anställda har avdelningen ”Executive” med 2st anställda.

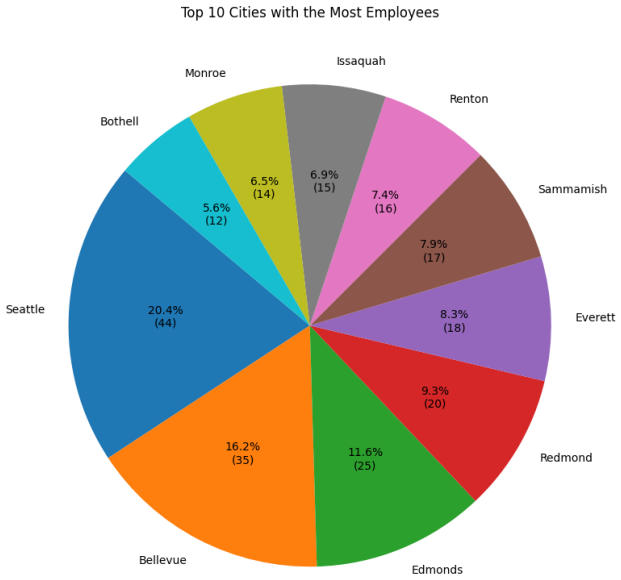


Könsfördelningen i AdventureWorks visar en tydlig dominans av män bland de anställa, där 71 % är män och endast 29 % är kvinnor. Detta kan bero på branschspecifika faktorer, såsom historiska normer eller arbetsuppgifter som traditionellt har lockat fler män. Samtidigt utgör detta en möjlighet att arbeta för en mer jämställd representation inom företaget. Att adressera dessa skillnader kan bidra till att skapa en mer inkluderande och diversifierad arbetsmiljö som gagnar både organisationen och dess medarbetare.



PERSON

Person schemat kopplar samman alla företagets affärsenheter med personer, deras kontaktuppgifter och roller. Diagrammet nedan illustrerar top 10 antalet anställda per stad i USA. Vi kan utläsa att flest anställda finns i Seattle med 44 personer med minst anställda finns i Bothell med 12 anställda.

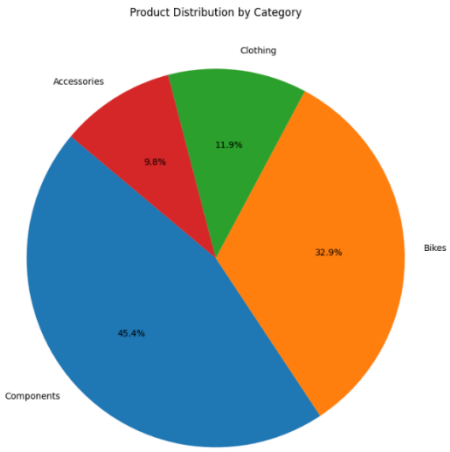


PRODUCTION

Som tidigare nämnt är AdventureWorks ett företag som fokuserar på att sälja och tillverka cyklar, komponenter och tillbehör. I databasen AdventureWorks2022 innehåller **Production-schemat** information om produkter och kategorier, exempelvis från tabeller som **Production.ProductCategory** och **Production.Product**.

Tabeller som **Production.ProductCategory** och **Production.Product** ger information om företagets produkter, uppdelade i kategorier som exempelvis:

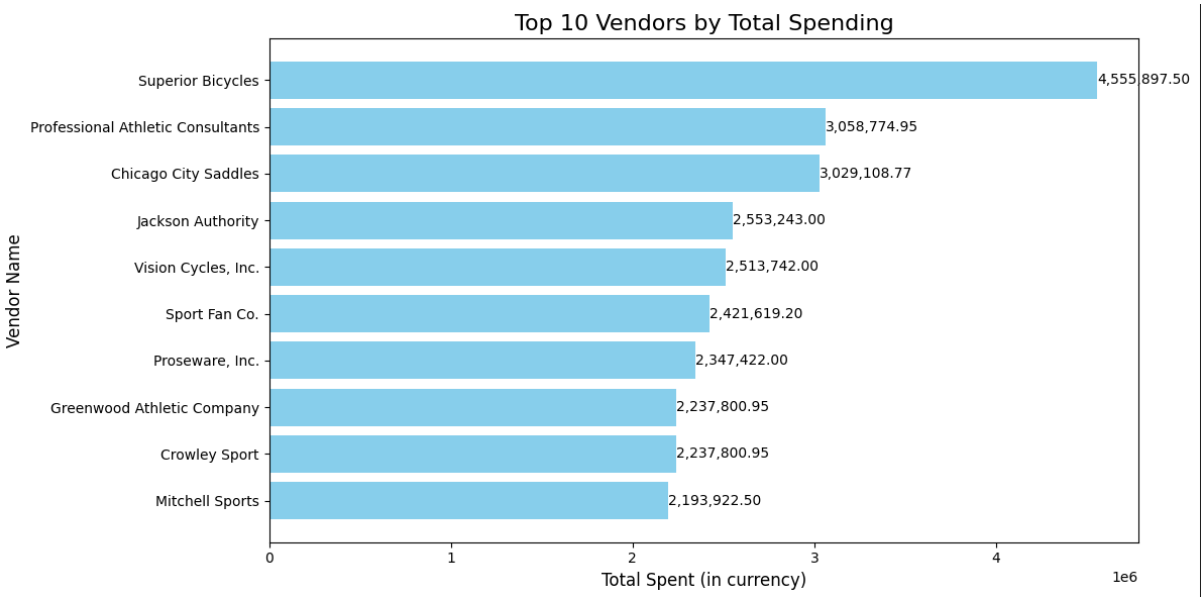
- **Komponenter**
- **Cyklar**
- **Kläder**
- **Accessoar**



CategoryName	ProductCount
Components	134
Bikes	97
Clothing	35
Accessories	29

PURCHASING

Purchasing schemat innehåller bland annat information om AdventureWorks återförsäljare, ordrar och leveranssätt. Tabellen nedan illustrerar en graf på de tio största återförsäljarna hos AdventureWorks, baserat på inköpssumma (USD).

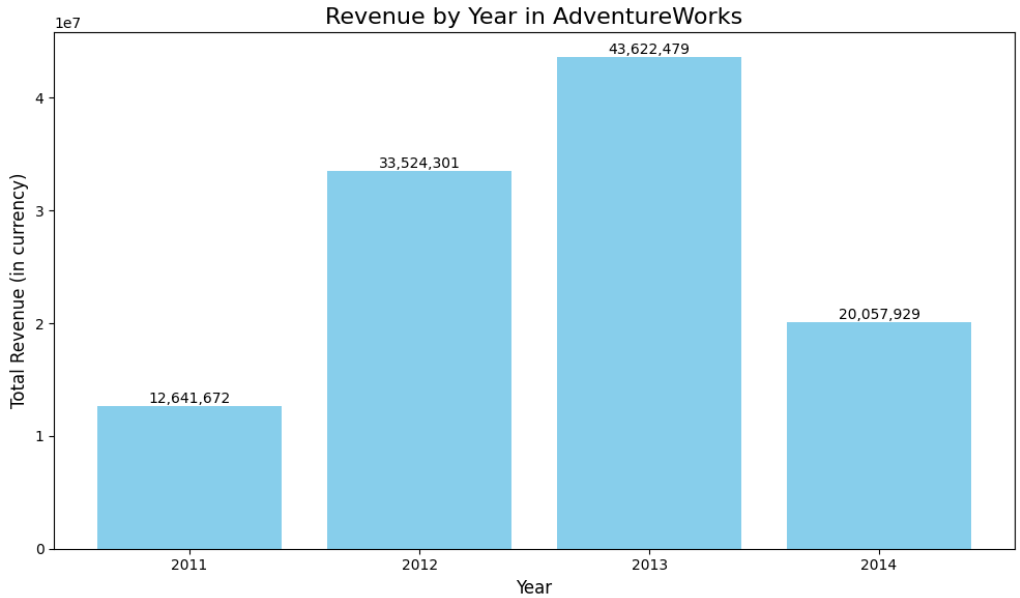


SALES

Sales schemat innehåller information om AdventureWorks försäljningar. Tabellerna i denna kategori innehåller bland annat information om kunder, valuta, säljare, säljskatt och omsättning. Tabellen nedan illustrerar företagets omsättning år 2011-2014. Ur den kan vi utläsa att mest omsättning hade företaget år 2013 (43,662,479 USD). Minst omsättning hade företaget år 2011 (12,641,672 USD).

AdventureWorks visar en årlig omsättning som sträcker sig från mer än 12 miljoner dollar till över 43 miljoner dollar. Datan visar en stadig ökning mellan åren 2011 och 2013. Denna ökning av omsättning kan indikera en stabil efterfrågan samt effektiv försäljningsstrategi. Under 2014 visas en kraftig minskning jämfört med föregående år. Denna nedgång beror dock inte på verklig minskning av försäljning utan snarare på ofullständig data från andra halvan av 2014, vilket ger en missvisande bild i diagrammet.

En annan tydlig förändring observeras mellan år 2011 och 2012, där omsättningen ser ut att öka mycket. Detta beror på att data endast finns tillgänglig från och med maj 2011, vilket gör att omsättningen för 2011 är baserad på ett ofullständigt år. Trots detta fortsätter trenden att visa en ökning mellan 2012 och 2013, vilket ger en indikation på företagets tillväxt och framgång under denna period.



Del två Analys delen

Gör en statistisk analys av valfri del av datan. Den skall innehålla åtminstone ett konfidensintervall. Hur tolkar du resultaten?

Den datan som är mest intressant för mig att undersöka och att analysera är AdventureWorks ekonomiska prestation med fokus på den årliga omsättningen. För att uppnå detta kommer jag att genomföra en statistisk analys med hjälp av både hypotesprövning och beräkning av ett konfidensintervall. Det jag vill komma fram till är att utvärdera stabiliteten i försäljningen över olika år och identifiera potentiella avvikelser som kan påverka företagets prestation.

Som vi kan se med hjälp av graferna ovan finns det skillnader i den årliga omsättningen mellan åren 2011 och 2014. Där år 2013 visar mest försäljning.

För att genomföra analysen så ställs följande hypoteser upp :

Nollhypotes (Ho) där $\mu_1 = \mu_2$, vilket innebär att den genomsnittliga omsättningen för två olika år är densamma.

Det ska också finnas en Alternativ hypotes (H1) där $\mu_1 > \mu_2$, vilket innebär att den genomsnittliga omsättningen för det ena året är högre än för det andra.

Utöver detta beräknas ett **95 % konfidensintervall** för den genomsnittliga årliga omsättningen. Konfidensintervallet används här för att uppskatta det sanna medelvärdet för omsättningen och dess osäkerhet.

Genomsnittlig årlig omsättning: 27,461,595.35

Standardavvikelse: 13,812,735.26

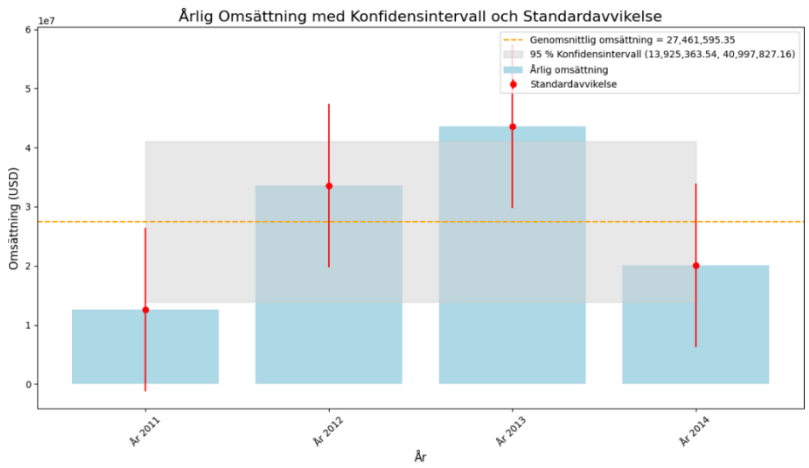
Antal år: (n) = 4

95 % Konfidensintervall: (5,482,451.21, 49,440,739.49)

Med hjälp av beräkning kommer vi fram till

Grafen visar:

1. **Årlig omsättning (ljusblå staplar):** Representerar omsättningen för varje år.
2. **Genomsnittlig omsättning (orange streckad linje):** Markerar medelvärdet av omsättningen, som är 27,461,595.35 USD.
3. **95% konfidensintervall (grått skuggat område):** Intervallet från 5,482,451.21 USD till 49,440,739.49 USD visar osäkerheten kring medelvärdet.



4. **Standardavvikelse (röd punkt och linje):** Den röda punkten markerar medelvärdet, och linjen visar standardavvikelsen (13,812,735.26 USD).

Graferna visar den årliga omsättningen för AdventureWorks mellan fyra år, där det genomsnittliga värdet är 27,461,595.35 USD. Standardavvikelsen på 13,812,735.26 USD indikerar en viss variation mellan åren, medan det breda 95% konfidensintervallet (5,482,451.21 - 49,440,739.49 USD) antyder osäkerhet kring det genomsnittliga värdet. Detta betyder att AdventureWorks har upplevt stora skillnader i sin ekonomiska prestation mellan åren, vilket bör analyseras vidare för att identifiera bakomliggande faktorer.

Standardavvikelsen är avgörande för att förstå spridningen i den årliga omsättningen, eftersom den visar hur mycket omsättningen varierar från genomsnittet. I vårt fall, med en standardavvikelse på 13,812,735.26 USD, framgår det att det finns en viss variation mellan de olika åren, vilket tyder på att omsättningen inte är helt stabil. Detta kan indikera påverkan från externa faktorer som marknadsförändringar eller interna strategier. För AdventureWorks är denna information viktig för att identifiera möjliga orsaker till variationen och för att utveckla strategier för att förbättra den ekonomiska stabiliteten framöver

Ordered Monthly Revenue Table (2011 - 2014):

SalesYear	2011	2012	2013	2014
SalesMonth				
Jan		3,970,627	2,087,872	4,289,818
Feb		1,475,427	2,316,922	1,337,725
Mar		2,975,748	3,412,069	7,217,531
Apr		1,634,601	2,532,266	1,797,174
May	503,806	3,074,603	3,245,624	5,366,675
Jun	458,911	4,099,354	5,081,069	49,006
Jul	2,044,600	3,417,954	4,896,354	
Aug	2,495,817	2,175,637	3,333,964	
Sep	502,074	3,454,152	4,532,909	
Oct	4,588,762	2,544,091	4,795,813	
Nov	737,840	1,872,702	3,312,130	
Dec	1,309,863	2,829,405	4,075,487	

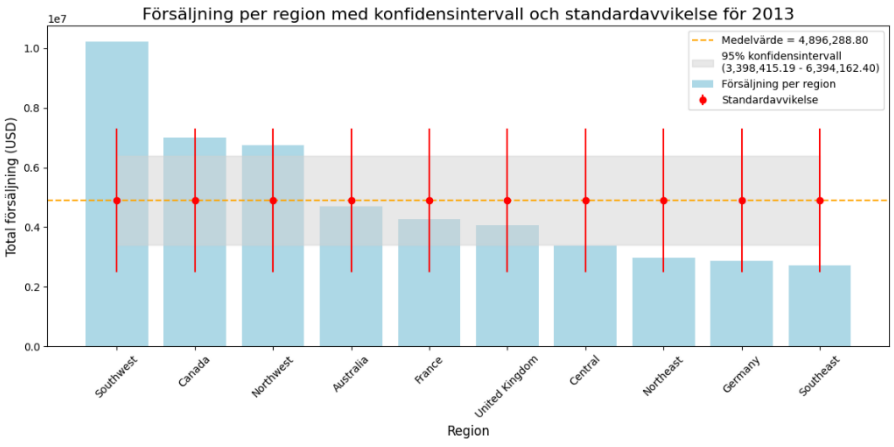
Nu analyseras år 2013 omsättning per region.

Genomsnittlig försäljning per region: 4,896,288.80 USD

Standardavvikelse: 2,416,724.12 USD

95 % konfidensintervall: 3,398,415.19 USD - 6,394,162.40 USD

Antal regioner (n) = 10



Staplar (ljusblå): Visar försäljning per region.

Medelvärde (orange linje): Representerar genomsnittlig försäljning över alla regioner.

Konfidensintervall (grått skuggat område): Anger intervallet där det genomsnittliga värdet sannolikt ligger med 95 % säkerhet.

Standardavvikelse (röda punkter och linjer): Visar hur mycket försäljningen varierar kring medelvärdet. Röda punkter markerar medelvärdet, och linjerna representerar standardavvikelsen.

Del 3

2. Analysera datan och resultaten du tagit fram. Vilka slutsatser och rekommendationer kan du ge?

AdventureWorks årliga ekonomiska prestation har här analyserats med fokus på omsättningen mellan åren 2011 och 2014 som datan visar. Denna period visar betydande variationer i omsättningen, och analysen syftar till att utvärdera stabiliteten, identifiera potentiella avvikelser och ge rekommendationer för att förbättra den ekonomiska utvecklingen. Genom att använda statistiska metoder, inklusive hypotesprövning och konfidensintervallberäkningar, kan vi skapa en djupare förståelse för företagets prestation under dessa år.

Den genomsnittliga omsättningen är **27,461,595.35 USD**. Det 95% konfidensintervallet för genomsnittlig omsättning är **13,925,363.54 - 40,997,827.16 USD**, vilket visar en stor osäkerhet. Standardavvikelsen är hög (**13,812,735.26 USD**), vilket tyder på stor variation i omsättningen mellan åren. År 2012 och 2013 sticker ut som år med högre omsättning jämfört med de andra åren.

Slutsats:

Omsättning är ej jämnt fördelan mellan åren. Det finns stora skillnader i omsättning mellan åren. År 2011 och år 2014 visar betydligt mindre omsättning än åren 2012 och 2013. Detta skulle i vanliga fall kunnat bero på externa faktorer såsom perioder med bättre marknadsförhållande eller marknadstrender. I detta fall beror den stora avvikelsen snarare på att företaget inte hade någon dokumenterad omsättning under de fyra första månaderna år 2011 samt under det andra halvåret för år 2014. Anledningen till att dessa månader ej har dokumenterad omsättning i databasen är okänd, men det kan spekuleras att företaget antingen helt slutade sälja produkter, eller att man inte har fortsatt att uppdatera databasen efter första halvåret 2014. Detta i sin tur gör det problematiskt att komma med konkreta rekommendationer för företagets fortsatta tillväxt. Jag har därför valt att analysera ytterliggare en aspekt av företaget, nämligen årlig omsättning per region.

Observationer: Försäljning per region med konfidensintervall och standardavvikelse för 2013

- Den genomsnittliga försäljningen per region är **4,896,288.80 USD**.
- 95% konfidensintervallet för genomsnittlig försäljning är **3,398,415.19 - 6,394,162.40 USD**.
- Regionen **Southwest** har markant högre försäljning (**10,239,209.34 USD**) jämfört med andra regioner, vilket gör den till en tydlig toppregion.
- Regionen **Southeast** har den lägsta försäljningen (**2,705,730.97 USD**), vilket kan indikera låg efterfrågan eller mindre marknadsnärvaro.

I detta fall har jag valt att endast analysera datan från år 2013 då omsättning var som högst och under teoretiskt förutsättning att företaget är fortsatt aktivt i att producera och sälja produkter.

Baserat på analysen kan vi se att Regionen SouthWest är en kritisk marknad för företaget och borde vara högprioriterat i framtiga strategier. De flesta andra regionerna/länderna ligger nära medelvärdet vilket tyder på en relativt jämn fördelning av försäljningen. Det finns dock potential att förbättra försäljningen regionalt i SouthEast och internationellt i Tyskland.

Rekommendationer:

Företaget bör analysera skillnaden i försäljning mellan regionerna. Framförallt borde man ta reda på orsaken till de relativt låga försäljningssiffrorna i regionen SouthEast och i Tyskland. Beror de låga siffrorna på mindre efterfrågan eller en mindre marknadsnärvaro från företagets sida? Genom att investera i kvalitativ datainsamling kan man analysera försäljningsdatan från de olika regionerna. Dessa kan inkludera demografisk information, kundbeteende och marknadstrender. Beroende på orsaken, kan man utveckla specifika strategier för att öka

försäljningen i dessa regioner. Man kan t.ex satsa på fler marknadsföringskampanjer med regionala och säsångsanpassade kampanjer eller öka fokus på lokala behov för att eventuellt införa nya produkter anpassade för just den specifika marknaden.

I övrigt borde man även utföra en djupanalys av de framgångsfaktorer som återfinns i SouthWest regionen och försöka överföra dessa strategier till de andra regionerna. Samtidigt borde man dra nytta av denna starka marknad och öka produktutbudet, till exempel genom introducering av premiumprodukter, för att maximera intäkterna från en redan framgångsrik marknad.

Executive summary

AdventureWorks är ett företag som fokuserar på cyklar och relaterade produkter. En analys av företagets databas, AdventureWorks2022, har flera insikter identifierats. Detta har identifierats genom nyckelinsikter kring personal, produktutbud och försäljningsprestationer. Företaget har totalt 296 anställda, varav 71 % är män, vilket indikerar en ojämn könsfördelning. Geografiskt är majoriteten av de anställda koncentrerade i Seattle. Flest anställda finns i Seattle, medan andra regioner har betydligt färre. Produktsortimentet domineras av komponenter (134 produkter), följt av cyklar och kläder.

Omsättningsdata för 2011–2014 visar en topp på 43,7 miljoner USD 2013 och ett genomsnitt på 27,5 miljoner USD, men med stora variationer mellan åren. År 2014 hade en nedgång på grund av ofullständig data, vilket begränsar analysens tillförlitlighet. En djupare analys av 2013 visar att Southwest presterar bäst (10,2 miljoner USD), medan Southeast ligger lägst (2,7 miljoner USD). Standardavvikelser och konfidensintervall indikerar att det finns osäkerheter och skillnader mellan regioner och år.

Rekommendationer inkluderar att förbättra datainsamlingen, analysera låga försäljningssiffror i specifika regioner och använda framgångsstrategier från Southwest för andra marknader. AdventureWorks bör även satsa på marknadsföring och produktanpassning för att öka tillväxten i underpresterande regioner. Detta skapar möjligheter att stärka företagets marknadsposition och ekonomiska stabilitet. Att utvärdera dessa områden kan bidra till förbättrad stabilitet och en hållbar affärsutveckling.

Självvärdering

Efter att du är klar skall du även skriva en kort redogörelse i slutet av rapporten där du

beskriver:

1. Utmaningar du haft under arbetet samt hur du hanterat dem.

Under arbetets gång stötte jag på flera utmaningar. Den första handlade om att förstå komplexiteten i databasstrukturen, eftersom AdventureWorks-databasen innehåller flera scheman och tabeller med olika relationer och data. Jag löste detta genom att noggrant studera databasens dokumentation och använda enkla SQL-frågor för att stegvis utforska tabellerna.Den andra utmaningen som jag stötte på var själva datakvaliteten. Med detta menar jag hanteringen av data som var ofullständig: Omsättningsdata för vissa år saknades, vilket påverkade analysens pålitlighet. Jag hanterade detta genom att noggrant dokumentera begränsningarna i rapporten och föreslå förbättrad datainsamling som en rekommendation.

2. Vilket betyg du anser att du skall ha och varför.

Jag anser att jag bör få betyget godkänt
Jag har genomfört en omfattande analys av AdventureWorks-databasen och använt både SQL och Python för att dra meningsfulla slutsatser. Min rapport innehåller en tydlig struktur, realistiska rekommendationer och välgrundade visualiseringar.Jag har hanterat utmaningar genom självstudier och löst problem på ett systematiskt sätt.

3. Tips du hade gett till dig själv i början av kursen nu när du slutfört den.

Definitivt tips jag har till mig själv är att börja med små frågor och bygga upp förståelsen steg för steg, istället för att försöka skriva avancerade frågor från början. Spendera mer tid med Datacamp-övningar, eftersom de hjälper väldigt mycket, och man kan faktiskt öva på det som man har läst i boken och gått igenom i klassen.