Teoretiska frågor

## 1. Beskriv kort hur en relationsdatabas fungerar.

En relationsdatabas har data i form av tabeller med rader och kolumner, där varje rad har en unik identifierare eller nyckel som kan koppla samman tabellerna för att upprätthålla en relation. Nyckeln används senare för att hitta relaterade data i relationsdatabasen.

## 2. Vad menas med ”CRUD” flödet?

CRUD är en akronym för Create, Read, Update, Delete, och det är flödet en databas bör gå igenom för att funka som en bra databas. Ny data ska skapas (Create), man ska kunna läsa datan (Read), man ska kunna uppdatera existerande data (update), och man ska kunna radera data från databasen (Delete). Detta är stöttepelare för en att en relationsdatabas ska vara användbar och va till någon nytta.

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

En innner join samlar ihop all data från två tabeller där värdena är detsamma från två utvalda kolumner från varsin tabell (ex. Samlar ihop all data där User\_id och Order\_id har samma värde).

Left join visar alla rader från första tabellen och de som matchar i den andra. Om det inte finns någon match mellan datan så lämnas ett NULL värde.

## 4. Beskriv kort vad indexering i SQL innebär.

En innner join samlar ihop all data från två tabeller där värdena är detsamma från två utvalda kolumner från varsinn tabell (ex. Samlar ihop all data där User\_id och Order\_id har samma värde).

Left join visar alla rader från första tabellen och de som matchar i den andra. Om det inte finns någon match mellan datan så lämnas ett NULL värde.

## 5. Beskriv kort vad en vy i SQL är.

En vy är en virtuell tabell gjord av en query som man kan framkalla med vyn istället för att behöva skriva om sin query en massa gånger. Det är ett enklare sätt att visa upprepande data som man vill komma åt ofta och fort.

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

En lagrad procedur är kod som kan sparas och bli återanvänd om och om igen. Man kan kalla på sin lagrade procedur för att framkalla den. Det kan också anta parametrar till skillnad från en vy men annars är de ganska lika. Stora skillnaden är att en vy är en virtuell tabell medans en lagrad procedur är spara kod som bli manipulerad.

Rapport

# Introduktion

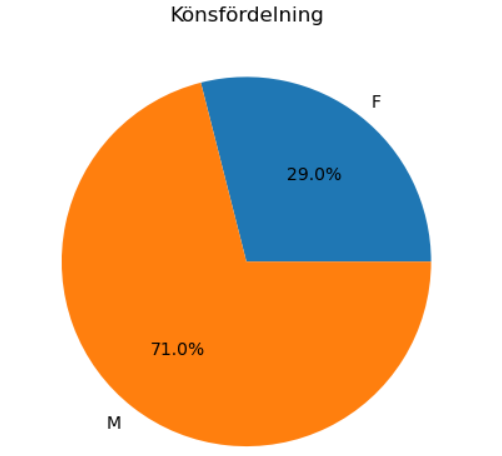
Denna rapport presenterar en analys av AdventureWorks2022databasen. Syftet är att ge en överblick över den data som finns och identifiera intressanta mönster och insikter i databasen.

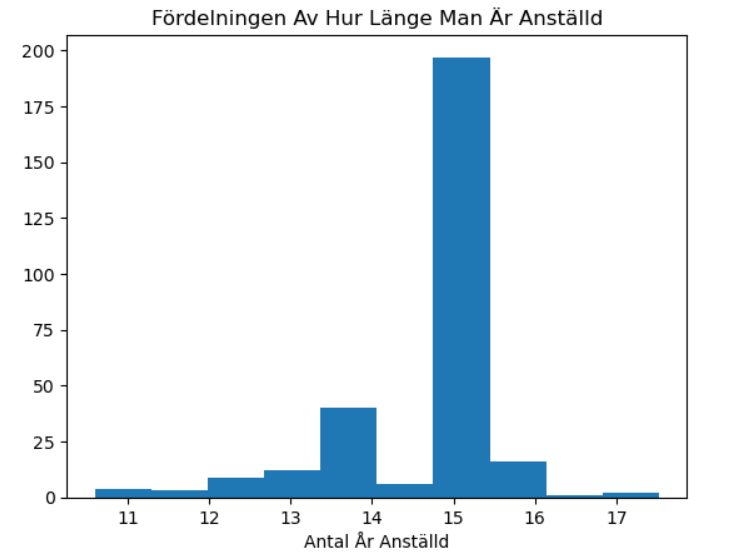
# Deskriptiv sammanfattning av databasen

AdventureWorks2022-databasen är en välstrukturerad databas som representerar olika aspekter av cykelföretagets verksamhet. Databasen innehåller flera tabeller. För denna analys kommer att fokusera på följande tabeller då de ger en bra överblick av företaget samt ger intressant data som är av intresse.

## HumanResourses.Employee

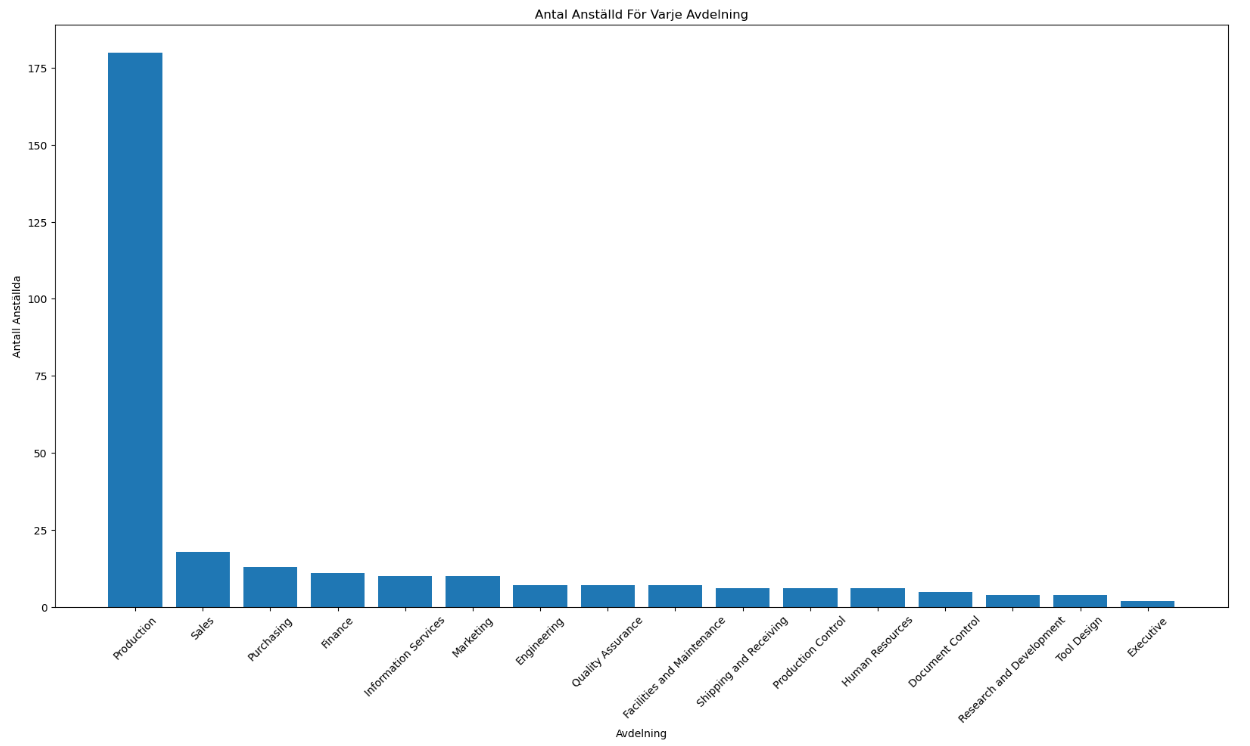
Denna tabell ger en omfattande överblick över AdventureWorks anställda, inklusive deras jobbtitlar, kön och annan information som är av intresse. Detta gör det möjligt att utföra intressanta analyser företagets personalstruktur. Vi kan se i datan att personalen är utgörs av 206 stycken män (71%) samt 84 kvinnor (29%). Att en person jobbar i snitt på AdventureWorks i ca 15 år och den som har jobbat där längst är grundaren som har jobbat där i ca 18 år, detta är givet att dem fortfarande jobbar där än idag (2023-12-27).





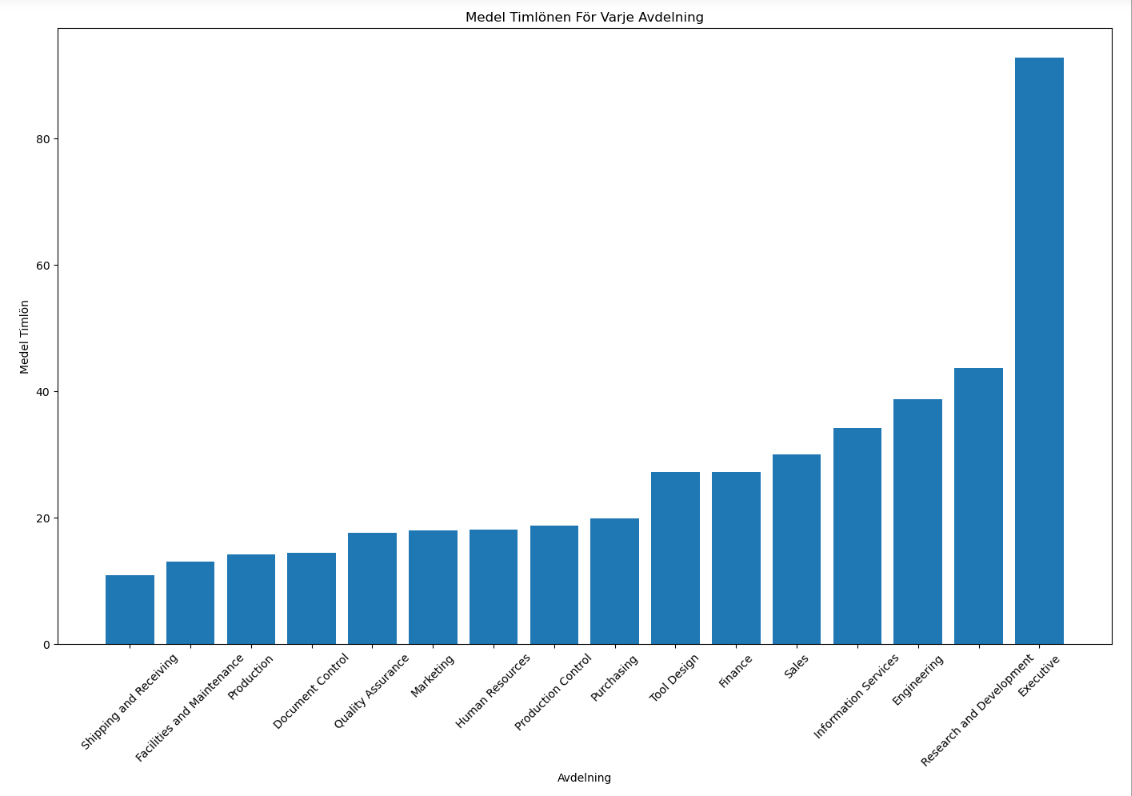
## HumanResourses.Department

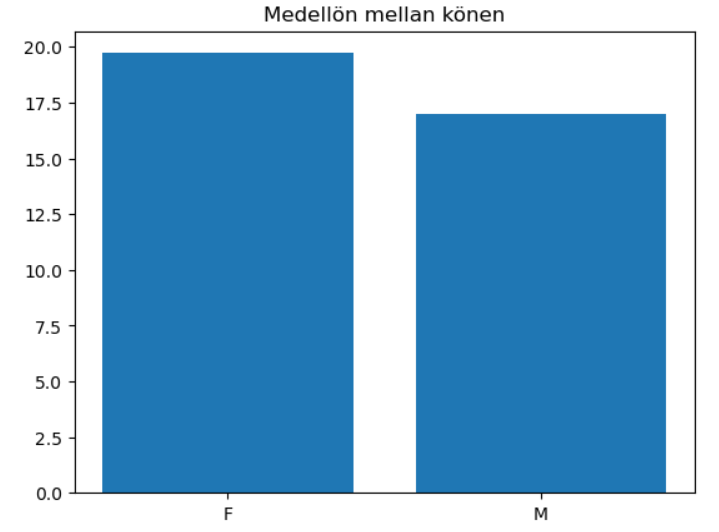
Denna tabell visar de 16 olika avdelningar som finns på AdventureWorks. Med hjälp av Employee tabellen kan vi analysera båda tabeller tillsammans och finna intressanta kopplingar. Vi kan se att den avdelningen med flest arbetare är i produktionsavdelningen med 180 arbetare, och avdelningen med minst arbetare bortsett från Executive avdelningen är Research and Development och Tool design med 4 arbetare var.



## HumanResourses.PayHistory

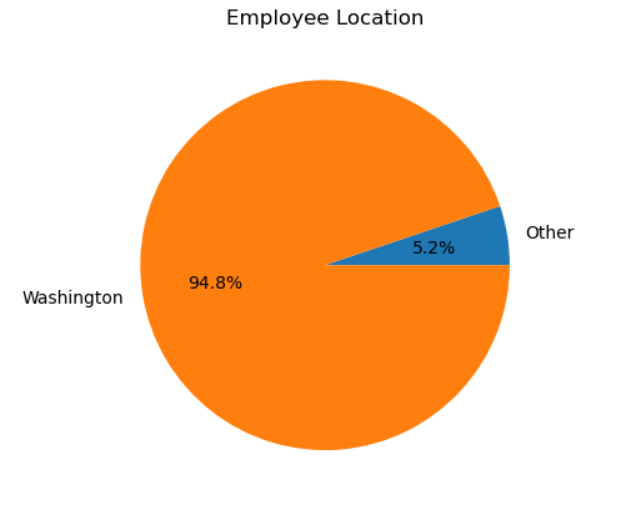
PayHistory visar alla anställdas timlön och hur den har ändrats över tid för de som har fått sin lön ändrad. Med hjälp av Department och Employee tabellerna kan vi få väldigt intressant data om vad de olika anställda får i timlön, samt kan vi se vilken avdelning som betalar mest på AdventureWorks. Vi ser att den med högst timlön är CEO:n för företaget och han tjänar 125,50 dollar per timme, medans den lägst betalda tjänar 6,50 dollar per timmen och har jobbtiteln Product Technichian - W30. Vi ser också att den högst betalda avdelningen i företaget är Executives med snittlönen 93 dollar per timme i snitt, därefter kommer Research And Development med snittlönen 43 dollar per timme. Den lägst betalda avdelningen är Shipping and Receiving med en snittlön på 11 dollar per timme. Något intressant i datan som vi kan se är att kvinnorna tjänar mer i medellön jämfört med männen, där kvinnorna tjänar ca 20 dollar per timme i medellön medan männen har 17 dollar per timme. Varför kvinnor har en högre medellön kan bero på att det jobbar fler män i produktion som tjänar väldigt låg lön och drar ner medellönen, medan kvinnorna jobbar i högre uppsatta jobb med högre lön.





## Person.BusinessEntityAdress

Person.BusinessEntityAddress är en tabell som innehåller information om adresser för personer och företag i AdventureWorks. Om vi tittar på antalet anställda i varje region, så kan vi se att Washington har flest anställda med **275** personer, följt av California med **2**. På andra sidan har Ontario (Canada), Oregon (USA) och Yukon Territory (Canada) endast **1** anställd vardera enligt tabellen. Detta ger oss en bra överblick över var AdventureWorks har sina anställda och vilka regioner som har flest respektive minst anställda.



## Production.Product

Product tabellen innehåller alla produkter som företaget säljer, hur mycket det kostar och mer detaljer om produkterna. Tabellen innehåller kolumner som produktID, Name med namnet på produkten, ReorderPoint med kvantiteten då de ska köpa in mer av produkten, när de började sälja varan och annan information som kan vara nödvändigt att veta om produkten. Tabellen innehåller 504 rader vilket motsvarar alla de olika produkterna som företaget eventuellt säljer eller har slutat sälja.

## Production.ProductInventory

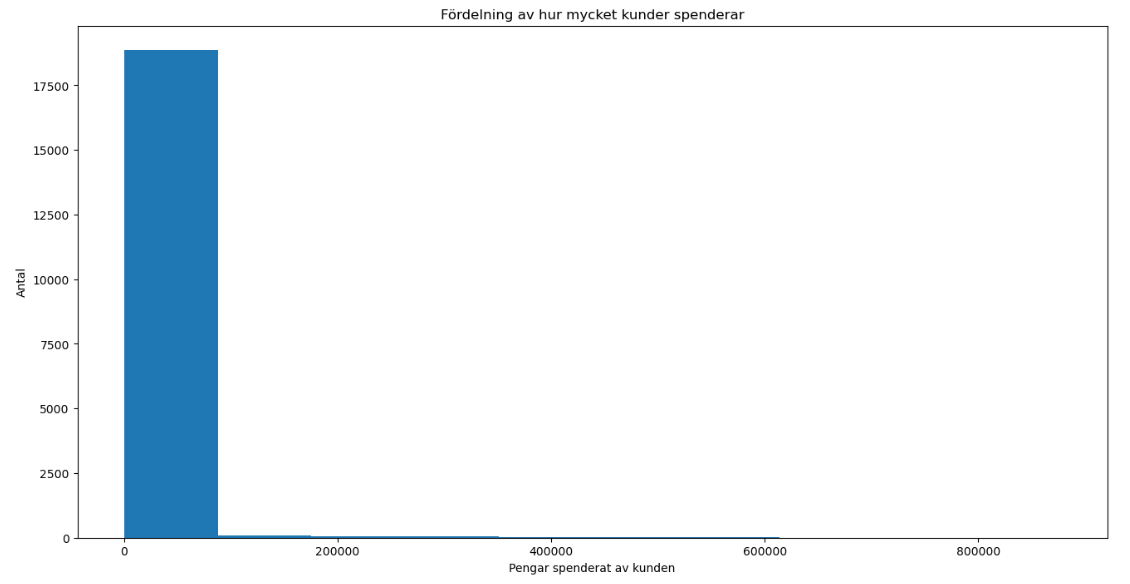
Tabellen visar vart alla produkter ligger i deras lager samt hur mycket som finns kvar av produkten. När vi kopplar ihop Product och ProductInventory tabellerna kan vi hitta de produkter som ligger under ReorderPoint. Dem produkter behöver fyllas på snarast för att inte de ska ta sluta i lager. Kollar vi då på den nya tabellen ser man att det är 7 olika produkter som behöver påfyllning i lagret medan resten har ligger över nivån för påfyllning.

## Production.ProuctListPriceHistory

Tabellen visar produkternas prisförändring över tiden. Tabellen innehåller kolumner som StartDate som visar när de börja med att sälja på det priset, EndDate som visar när de sluta med priset och ListPrice som antingen visar nuvarande pris eller förgående pris beroende på vilket priset produkten för nuvarande innehar.

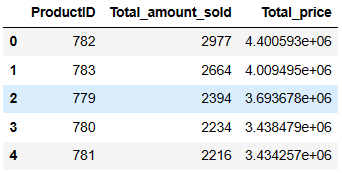
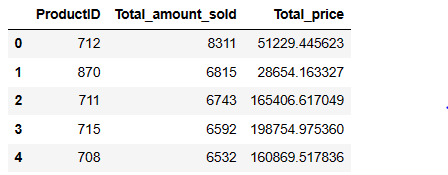
## Sales.Customer

Här ser vi kunderna som har köpt eller köper från oss. Tabellen visar vem kunden är, om det är en affär eller en person som köpt något och vart de har köpt från. Tabellen består av 19 820 rader vilket då står för 19 820 olika kunder. Ut av alla dessa kunder kan filtrera ut vilka som är personer eller en affär som är en kund och då ser man att där är 701 affärer som är kunder och att 18 484 är individuella personer som är kunder för AdventureWorks. Som vi kan se så är det betydligt mycket fler personer som köper från AdventureWorks än vad butiker köper från dem.



## Sales.SalesOrderDetail

Detta är en tabell som visar vilken har köpts för vilken order, hur mycket av varan som har köpts på ordern, vart produkterna ska levereras och hur mycket produkten kostar samt kostnaden för kvantiteten av produkten. Med denna tabell kan vi se vilka produkter som säljs mest av och vilka produkter som har get mest vinst. I tabellen ser vi att produkt med ID 712 har sålt mest med 8311 enheter men produkten är långt från den produkt som har genererat mest vinst för företaget. Den produkten har ID:et 782 och har generat cirka 4 miljoner dollar men som har sålts mindre än hälften än den som har sålts mest.



## Sales.SalesOrderHeader

Denna tabell är en utökning av SalesOrderDetail med ännu mer information om orderbekräftelsen. Den visar information som när ordern blev tillagd, när den blev skickad, vilken kund som har lagt order, leveransadress och olika kostnader så som frakt, skatt och totala summan för ordern. Detta är vad som behövs för en fullständig orderbekräftelse som visar allt det viktiga för en order. Med denna tabell och SalesOrderDetail kan vi ta fram de kunderna som spenderar mest och de som spenderar minst. Vi kan filtrera tabellen och se att den kund som spenderat mest har hittills spenderat 877 107 dollar på företaget vilket är ett väldigt högt värde då medelvärdet på spenderandet av kunder ligger på 3491. Dollar. Detta kan bero på att de kunder som spenderar mer är butiker som köper mer i bulk jämfört med personer som köper i mindre kvantitet, och eftersom det är betydligt fler personer som köper av företaget än butiker så kommer detta värde jämfört med medelvärdet att se betydligt mycket större ut än vad det egentligen är. För att få en tydligare bild får man använda butikerna som en enskild population i ett stickprov för att verkligen se hur butikerna spenderar i snitt och samma för personerna.

## Sales.SalesPerson

Detta är en tabell för de som jobbar inom försäljning. Den visar deras Commission, deras säljs-mål, hur mycket de har tjänat föra och under året. Tabellen har 17 rader som representerar de 17 olika säljare. Vi kan filtrera ut data från tabellen och se vilka som har klarat målet och vilka som säljer bäst för företaget. Vi finner då att alla säljare klarar målet och att de säljare som säljer bäst säljer för mer än 5 miljoner dollar.

# Statistisk analys

## Syftet med analysen och de frågor som ska besvaras.

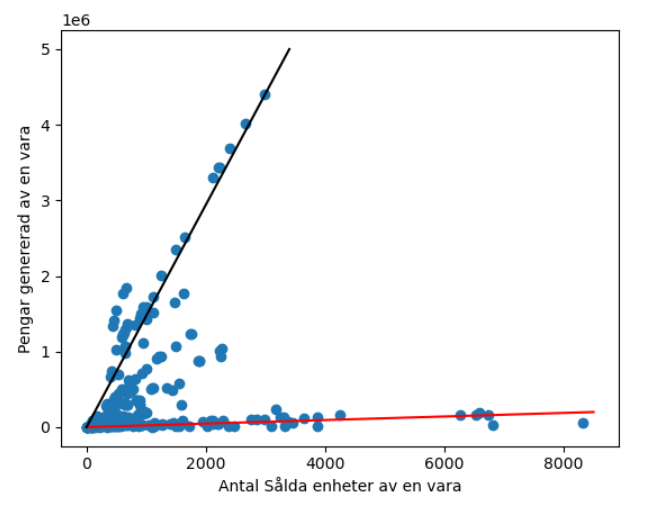
Syftet med denna analys är att undersöka försäljningsdata som finns i AdventureWorks2022 databasen och hitta intressanta aspekter i datan och möjligtvis besvara intressanta frågor som kan eventuellt hjälpa företaget tjäna mer pengar. All data kommer från AdventureWorks2022 databasen som tar data från ett cykelföretag. Frågor som kommer att besvara är: Vilka varor säljer mest?

## Statistiska metoder som har använts för att analysera all data.

För att analysera datan så kommer vanliga statistiska metoder att använda så som konfidensintervall, korrelationskoefficienter, standardavvikelse m.m. För att få fram datan ur databasen kommer jag att använda SQL-Server Managment Studios för att skapa querys för att framkalla tabeller av intresse, sedan skickar vi vidare all data in till python för att visualisera datan med hjälp olika grafer och för att räkna ut konfidensintervall och andra statistiska beräkningar. I python kommer olika libraries att användas för att förenkla processen. All kod finns på Github för all data som är framtagen i denna rapport. ([THEXBOXNERD81/SQL (github.com)](https://github.com/THEXBOXNERD81/SQL/tree/master))

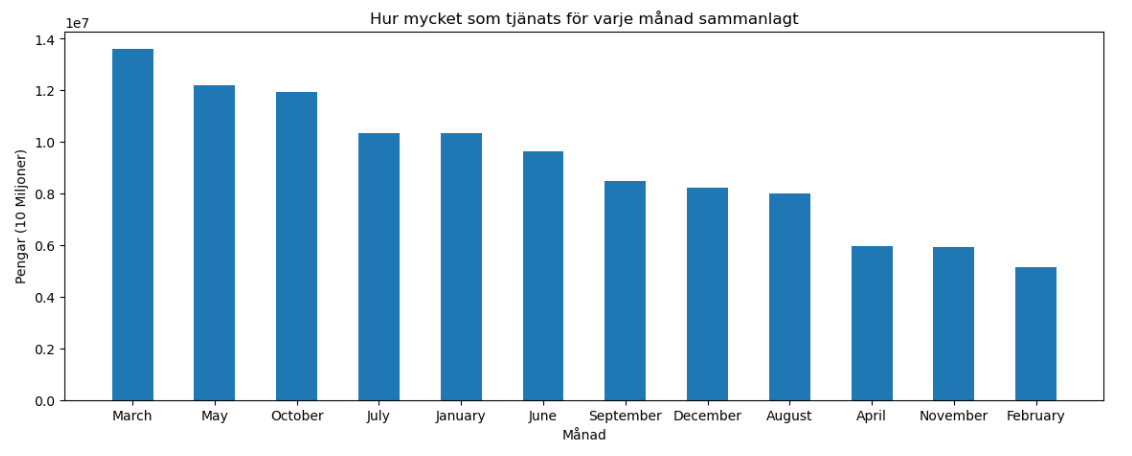
## Resultaten av analysen, inklusive tabeller, diagram och andra visuella hjälpmedel samt Diskussion av resultaten och vad de betyder för den fråga som undersöks

I den insamlade datan hittar vi i tabellen med varor ProductID, kvantitet sålt och Totalt intjänat vilka varor som säljer bäst. Vi ser i datan att de inte är de varorna som generar mest pengar för företaget. Om vi väljer att ordna datan enligt pris så får vi helt olika varor som säljer mindre än hälften så bra. För att undersöka detta vidare användes en scatterplot för att se om någon trend utformas av värdena mellan antal sålda varor och intjänade pengar från en vara.

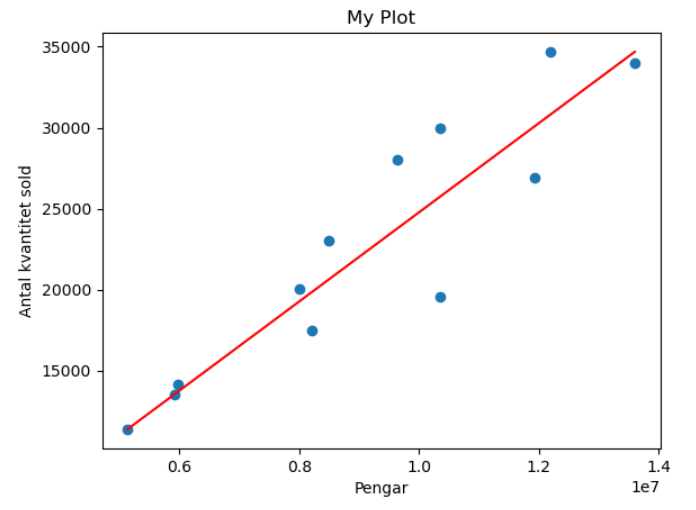


Korrelationskoefficienten för dessa två variabler ger värdet 0,15 vilket kan tolkas som att det knappt finns någon korrelation mellan variablerna, men kollar vi på grafen så kan vi hitta klara trender i datan. Vi kan utforma en trendlinje som ligger vid ungefär vid 0 och en annan som går väldigt högt upp. Den röda trendlinjen visar att där finns varor som det inte ger mycket vinst även om det säljs mycket av produkten. Vi kan se att den produkt som det sålts mest av har inte genererat mycket pengar för företaget. Den svarta trendlinjen visar tvärt om, en relativt låg försäljning med en hög vinst från varorna som säljs.

Går vidare och kollar på tabellen med försäljningsdata för varje månad kan vi hitta dem månader som säljer bäst, får in mest orders och som säljer mest kvantitet av varor. Vi kan i tabellen se att Mars månad generar mest pengar men att Maj säljer mest kvantitet av varor och får flest antal orders från kunder.

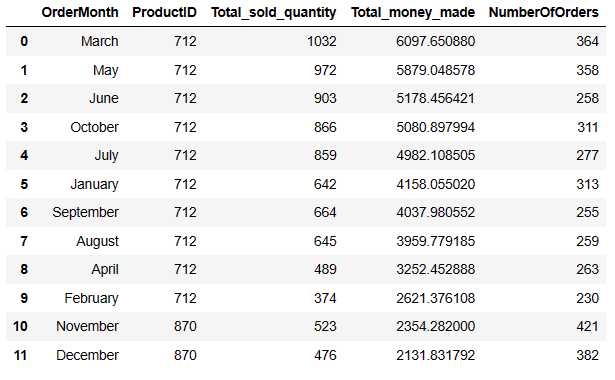


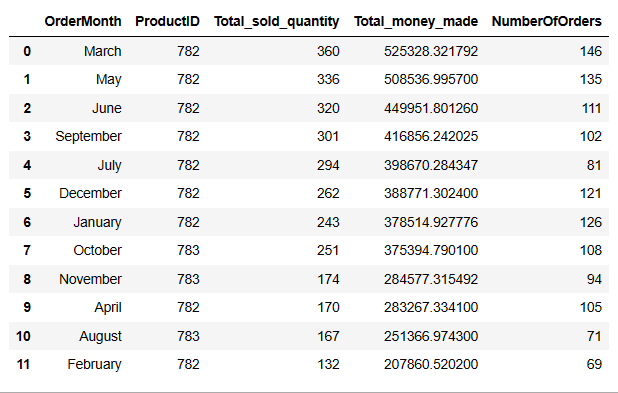
Gör vi en scatterplot för totalt intjänat och total kvantitet såld så ser man att där finns en klar korrelation mellan de två med hjälp av trendlinjen. Korrelationskoefficienten för dessa två variabler är 0,91 vilket kan tolkas som en väldigt tydlig positiv korrelation mellan de två variablerna. Detta förklarar också varför vi kunde hitta klara trendlinjer i den förra grafen då korrelationen mellan variablerna är väldigt tydlig. Varför den förra grafen inte gav någon bra värde på koefficienten kan bero på att de olika produkterna säljer olika bra och kostar olika mycket vilket kan ge olika trender beroende på hur nära varorna ligger i värde.



Kollar vi på konfidensintervallet för summan som man sammanlagt tjänat in för varje månad med Konfidensgraden 95% kan vi se om det finns några månader som sticker utanför konfidensintervallet. Ett värde som faller utanför konfidensintervallet tyder på att det är mer sannolikt att andra faktorer påverkar värdet än för de som ligger i konfidensintervallet. Vi ser då i tabellen att april, november och februari ligger under intervallet och oktober, mars och maj ligger över intervallet. Faktorer som kan ligga bakom detta kan bero till exempel på att det är utanför säsongen för att cykla och då köper man inte lika mycket cykel produkter som när det är säsong. Det kan också bero på en bra marknadsföring under de bra månaderna. För att ta reda på det måste vi dyka djupare in i databasen och försöka hitta paralleller mellan dem.

Vi kan också koppla ihop dessa tabeller för att se vilken produkt som säljer bäst under vilken månad. Här ser vi att produkterna med ID 712 och 870 säljer mest av under månaderna, vilket vi såg i tabell innan att dem produkterna sålde bäst. Den produkt som de tjänar mest på för månaderna är produkt 782 och 783, vilket vi också såg i förgående tabell. Varför produkterna inte kommer på varandras tabeller beror på den första scatterplotten vi såg från början. Vi ser där produkterna som säljs mycket av men som inte ger mycket vinst och sen har vi motsatsen.





## Slutsats

Slutsatsen som vi kan dra från vår data som är framtagen är att den produkt som säljs bäst av företaget är det med produkt ID 712, dock är det inte den varan som generar mest vinst för företaget, det är produkten med produktID 782 som tar det priset. Tittar man noggrannare i vår data så kan vi se för vilken månad som en produkt säljer som bäst. I dem tabellerna kan vi se att produkt 782 ger mest vinst för de flesta månaderna förutom i oktober, november och augusti då produkt 783 ger mest vinst.

Något som AdventureWorks borde kolla in i mer är deras försäljning och se vad som eventuellt påverkar deras vinster under de bra månaderna och de dåliga för att utveckla företaget och vinsterna. De borde också titta över produkterna som de säljer väldigt mycket av och inte tjänar så mycket på och eventuellt se över om de är värt att ha kvar dem, Samt att de borde fokusera mer på produkterna som de tjänar mycket på för att utöka vinsterna för företaget.

## Begränsningar

Eventuella begränsningar på analysen kan vara så som bortfall i data som kan ge ett snedvridet resultat, felmätningar som att data i databasen är inmatat på fel sätt eller har givit ett fel värde, stickprovet storlek kan ha varit för litet för vissa statistiska slutsatser, samt tester förlitar sig på antaganden och om de blir brutna så kommer resultaten att bli opålitligt. En till begränsning kan vara brist på förståelse, Det kan ha uppkommit fel antagande om kopplingar i vår data som har tolkats fel som gör att resultaten tolkas på ett annat sätt än vad verkligheten egentligen säger.

# Executive summary

Denna rapport presenterar en detaljerad analys av AdventureWorks2022-databasen, fokuserad på personalstruktur, löner, produktförsäljning och försäljningsdata. Genom att granska tabeller som Employee, Department och Sales, identifierar rapporten mönster och ger insikter om företagets verksamhet. Intressanta observationer inkluderar att kvinnor har en högre genomsnittslön jämfört med män och att vissa produkter, särskilt med produktID 782, genererar betydande vinster. Analysen av månadsvis försäljningsdata avslöjar starka korrelationer mellan intjänade pengar och såld kvantitet, medan konfidensintervall visar variation i försäljningen över månader. Slutsatserna betonar vikten av att förstå varför vissa produkter säljer bättre och strategier för att maximera vinsten baserat på försäljningstrender.

#### Muntlig presentation gjordes 2024-01-02

Reflektion på eget arbete

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

Den största utmaningen med arbetet var att skriva själva rapporten. Att försöka strukturera den och formulera sig på ett någorlunda vetenskapligt sätt har varit utmanande samt vad man egentligen ska skriva och inte skriva i rapporten. En annan utmaning var att veta vad som egentligen är intressant data och ointressant data. Då detta är en subjektiv fråga kan det bli svårt att hitta det man egentligen söker och tar fram ganska så ytlig data som kanske inte intresserar så mycket.

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

Stora delar av rapporten tycker jag faller under VG nivå exemplet i kunskapskontrolls pdf:en. Jag har till stora delar utforskat datan längre och försökt hitta intressanta kopplingar mellan olika tabeller samt tagit fram grafer som förenklar och visar mer tydligt resultaten som är framkallade under sammanfattningen och i analysen. Samt har flera statistiska metoder använts för att kunna hitta kopplingar i databasen och mellan tabellerna.

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

Att verkligen ta fram så specifik data som möjligt så det verkligen är rätt data som är framtagen, samt att försöka hitta en mall eller skapa en struktur för rapporten så fort som möjligt så att man får en över blick om hur och vad man ska skriva.