Kunskapskontroll: Siarhei Thor

# Innehållsförteckning

[**Innehållsförteckning 1**](#_k3ajfviv655a)

[**Teoretiska frågor. 2**](#_dxp5lsv3t2kx)

[**Programmeringsuppgift och Rapport 3**](#_rz0z5ebeo4vc)

[En deskriptiv sammanfattning över databasen AdventureWorks2022 3](#_ld1mj3iffc2z)

[Göra en statistisk analys. Vilka könsskillnader finns på företaget. 5](#_qoph9osws0ks)

[Sammanfattning av statistisk analys: Könsskillnader på företaget. 7](#_zi86okpyxptr)

[Övriga explorationer. 8](#_a586qw1m03e3)

# Teoretiska frågor.

1. Relationsdatabas:

En relationsdatabas är en typ av databas där data organiseras i tabeller, som kallas relationer. Varje tabell består av rader och kolumner, där varje rad representerar en datapost och varje kolumn en specifik attributtyp. Tabellerna kan vara länkade till varandra genom nycklar (primärnycklar och främmande nycklar), vilket möjliggör effektiv dataåterhämtning och integritet. Relationella databaser använder Structured Query Language (SQL) för att hantera och hämta data.

1. CRUD-flöde:

CRUD står för Create, Read, Update, Delete. Det är ett vanligt mönster som beskriver de grundläggande operationerna som utförs i databasapplikationer.

Create: Skapa nya poster i databasen.

Read: Läsa eller hämta befintliga data.

Update: Modifiera eller uppdatera befintliga data.

Delete: Ta bort data.

1. Left Join och Inner Join:

Left Join: Kombinerar data från två tabeller. Det returnerar alla rader från vänster tabell och matchade rader från höger tabell. Om det inte finns någon matchning, fylls resultatet med NULL på sidan av den högra tabellen.

Inner Join: Kombinerar rader från två tabeller baserat på en matchande villkor. Endast de rader som uppfyller villkoret från båda tabellerna inkluderas i resultatet.

Användningen av dessa join-typer beror på vilken data som behövs och relationen mellan tabellerna.

1. Indexering i SQL:

Indexering i SQL är en prestandaförbättringsteknik där en särskild datastruktur (index) skapas för att snabba upp dataåterhämtning. Index kan ses som ett innehållsförteckning för databasen som gör det snabbare att hitta specifika rader i tabeller baserat på indexkolumnerna. De är särskilt användbara vid sökning, sortering och filtrering av stora datamängder.

1. Vy i SQL:

En vy i SQL är en virtuell tabell som representeras av en SQL-fråga. Den innehåller inga data själv utan är ett lager som ger en anpassad vy över databasens data. Användare kan interagera med en vy precis som med en vanlig tabell, men vyens data genereras dynamiskt vid förfrågan från underliggande tabeller.

1. Lagrad procedur i SQL:

En lagrad procedur är en samling fördefinierade SQL-instruktioner som lagras i databasen. De används för att kapsla in ofta använda eller komplexa operationer så att de kan återanvändas och köras med en enda anrop. Lagrade procedurer kan förbättra prestanda, främja återanvändning av kod, och bidra till bättre säkerhet och underhåll av databasapplikationer.

# Programmeringsuppgift och Rapport

## En deskriptiv sammanfattning över databasen AdventureWorks2022

AW2022 är en relationsdatabasen och innehåler följande Schemas:  
**'dbo', 'HumanResources', 'Person', 'Production', 'Purchasing', 'Sales'**

Ett exetmpel på van HumanResources schema innehåller:

Schema Table

HumanResources Department

HumanResources Employee

HumanResources EmployeeDepartmentHistory

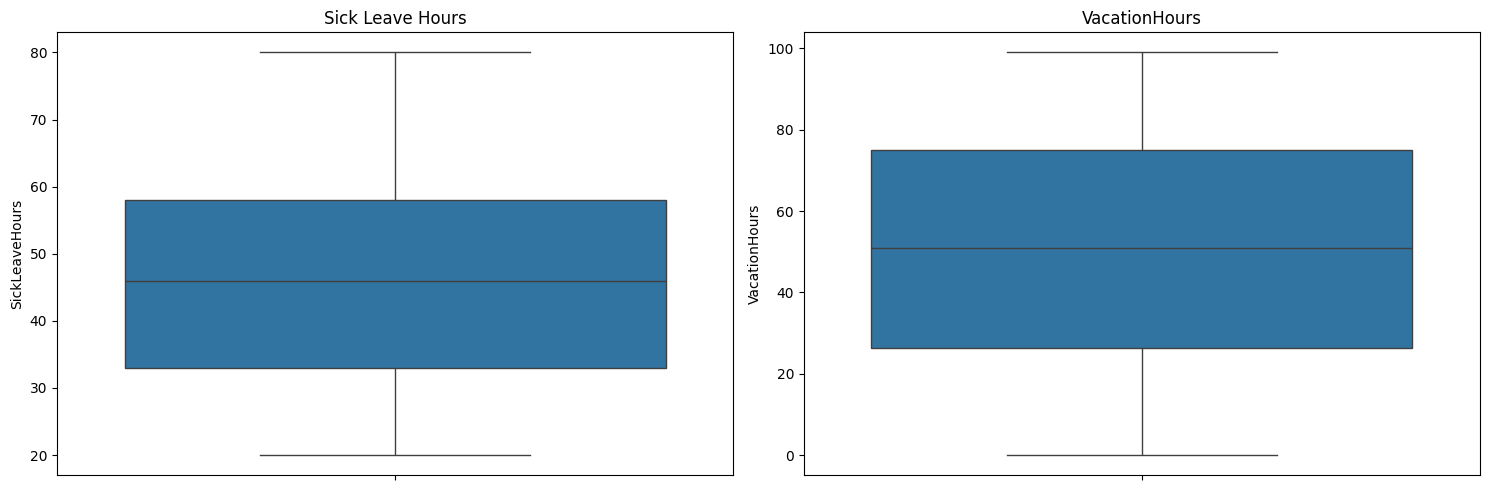
HumanResources EmployeePayHistory

HumanResources JobCandidate

HumanResources Shift

Descriptive Stats för Employees Table:

|  | **BusinessEntityID** | **OrganizationLevel** | **VacationHours** | **SickLeaveHours** |
| --- | --- | --- | --- | --- |
| count | 290.000000 | 289.000000 | 290.000000 | 290.000000 |
| mean | 145.500000 | 3.522491 | 50.613793 | 45.306897 |
| min | 1.000000 | 1.000000 | 0.000000 | 20.000000 |
| 25% | 73.250000 | 3.000000 | 26.250000 | 33.000000 |
| 50% | 145.500000 | 4.000000 | 51.000000 | 46.000000 |
| 75% | 217.750000 | 4.000000 | 75.000000 | 58.000000 |
| max | 290.000000 | 4.000000 | 99.000000 | 80.000000 |
| std | 83.860002 | 0.750240 | 28.786215 | 14.540444 |



Förtydligande:

**VacationHours**:

Denna variabel, med 290 observationer, har värden som sträcker sig från 0 till 99 timmar.

Den genomsnittliga semestertiden är cirka 50,61 timmar, med ett medianvärde nästan lika med medelvärdet (51), vilket indikerar en ganska symmetrisk fördelning.

Standardavvikelsen är 28,79, vilket visar på en varierad mängd semestertimmar bland enheterna.

**SickLeaveHours**:

Även med 290 observationer varierar sjuktimmar från 20 till 80 timmar.

Genomsnittet är ungefär 45,31 timmar, nära anknutet till medianvärdet på 46, vilket antyder en symmetrisk fördelning.

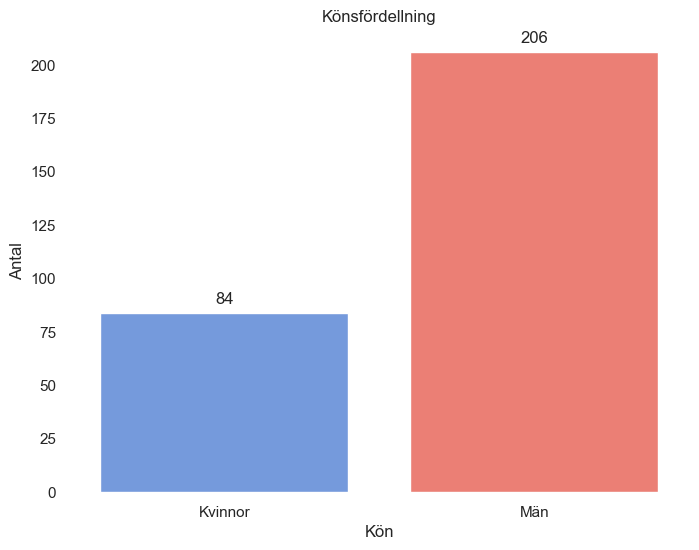
Standardavvikelsen är 14,54, mindre varierad jämfört med semestertimmar men fortfarande betydande.

## Göra en statistisk analys. Vilka könskillnader finns på företaget.

1. Antal anställda per kön.

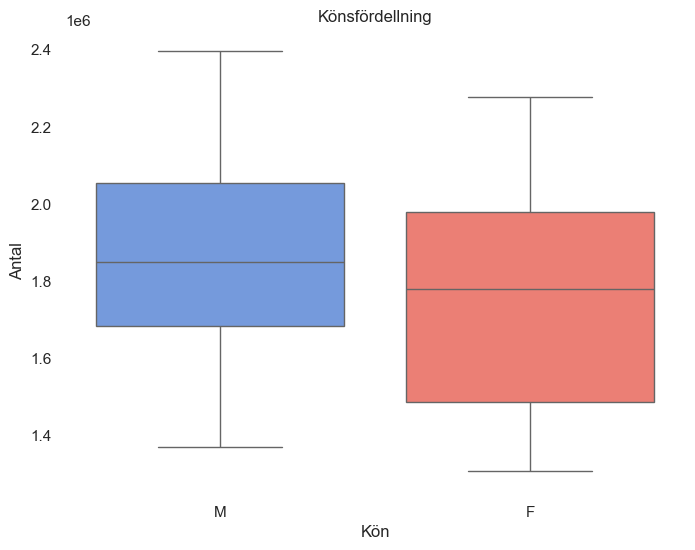
2. Finns det skillnad i total försäljning för män och kvinnor?

Antal anställda per kön. Kvinnor 84, Män 206.



Efter vi rensar data vi ser att total försäljning fördelat mellan män och kvinnor ser vi att män verkar känna in mer än kvinnor:

|  | **mean** | **max** | **min** |
| --- | --- | --- | --- |
| **Gender** |  |  |  |
| F | 1.764287e+06 | 2.278549e+06 | 1.307950e+06 |
| M | 1.871463e+06 | 2.396540e+06 | 1.371635e+06 |

Detta visar på en skillnad i genomsnitt -107176.0.   
  


Hypotesprovning.

H0 Nollhypotes: Det finns inget könsbaserad skillnad i försäljningen på AdventureWorks.

H1 Alternativ hypotes: Det finns en statistisk signifikant skillnad.

Signifikans nivå: 0.05.

T-statistic: -0.5459954290469845

95% Confidence Interval: (-539215.6789327734, 324864.476647059)

**P-value: 0.5966977048710802 - *H*0 *inte kan förkastas. Dvs det finns inget statistiskt signifikant könsbaserad skillnad i totala försäljningen på Adventure works.***

## 

## Sammanfattning av statistisk analys: Könsskillnader på företaget.

I analysen av AdventureWorks har vi undersökt antalet anställda per kön och eventuella skillnader i försäljning mellan män och kvinnor. Företaget har 290 anställda, varav 84 är kvinnor och 206 är män. I försäljningsdata visar män något högre genomsnittliga siffror än kvinnor.

En hypotesprövning genomfördes för att fastställa om dessa skillnader är statistiskt signifikanta. Med en signifikansnivå på 0.05 visade testet att nollhypotesen inte kunde förkastas (T-statistik: -0.546, P-värde: 0.597), vilket betyder att det inte finns några statistiskt signifikanta könsbaserade skillnader i försäljningen på företaget.

## 

## Övriga explorationer.

Jag har även kollat på försäljningen fördelat på området och produkter inkl datumen, SQL koden är lite mer avancerad, men detta faller utanför kunskap kontrollens scope. Ett exempel är att jag har gjort ett ANOVA test på ett se om det finns ett samband i tandutvecklingen mellan område och försäljning.

ANOVA

sum\_sq df F PR(>F)

C(TerritoryName) 2.866122e+09 9.0 57.226646 1.772187e-100

Residual 3.120214e+10 5607.0 NaN NaN

Låg PR(>F) tyder på statistisk signifikant samband. Då kan vi kola på Pairwise Tukey, för att se vilka har samband (skiljer sig signifikant).

quad\_r = quad(f, low, high, args=args, full\_output=self.full\_output,

Multiple Comparison of Means - Tukey HSD, FWER=0.05

=======================================================================

group1 group2 meandiff p-adj lower upper reject

-----------------------------------------------------------------------

Australia Canada 1658.8082 0.0 1278.9836 2038.6327 True

Australia Central 2594.8141 0.0 2049.6631 3139.9652 True

Australia France 808.6196 0.0 372.5246 1244.7145 True

Australia Germany 185.2416 0.9723 -302.7811 673.2643 False

Australia Northeast 2497.6646 0.0 1923.891 3071.4382 True

Australia Northwest 1672.6686 0.0 1304.9007 2040.4365 True

Australia Southeast 1585.2932 0.0 1061.434 2109.1523 True

Australia Southwest 1049.4381 0.0 708.356 1390.5202 True

Australia United Kingdom 830.0672 0.0 415.8436 1244.2908 True

Canada Central 936.006 0.0 384.6544 1487.3575 True

Canada France -850.1886 0.0 -1294.0103 -406.3669 True

Canada Germany -1473.5666 0.0 -1968.506 -978.6272 True

Canada Northeast 838.8564 0.0002 259.1884 1418.5245 True

Canada Northwest 13.8604 1.0 -363.0376 390.7585 False

Canada Southeast -73.515 1.0 -603.8237 456.7937 False

…..

Där vi har False, innebär att det inte finns någon statistisk skillnad och då kan vi evt gruppera dessa för en viss business relaterad ändamål.