Kunskapskontroll

Datum: 2024-01-05

Författare: Keikiet Pham

Kurs: SQL (25Yhp)

Lärare: Márk Mészáros, Antonio Prgomet

Muntlig Presentation: 2024-01-03

Del 1 - Teoretiska Frågor

1. Beskriv kort hur en relationsdatabas fungerar.

Relationsdatabaser är en typ av Database Management Systems (DBMS) som organiserar data i tabeller med rader och kolumner. I relationsdatabaser skapas relationer mellan två eller flera tabeller med hjälp av s.k. primary keys (pk) som kopplas till foreign keys (fk) i en annan tabell.

2. Vad menas med ”CRUD” flödet?

CRUD står för Create, Read, Update och Delete och är grundläggande operationer som används för att hantera data i olika databashanteringsystem där:

- *Create: Skapar nya poster i databasen.*
- *Read: Hämtar och läser data från databasen.*
- *Update: Uppdaterar befintliga poster i databasen.*
- *Delete: Tar bort poster från databasen.*

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

"Left join" och "Inner join" är vanliga "join" operationer i bl.a. SQL och används för att returnerna/hämta data från en annan tabell.

- *Left Join: Returnerar alla rader från den vänstra tabellen och matchande rader från den högra tabellen. Om det inte finns någon matchning i högra tabellen, returneras NULL-värden.*
- *Inner Join: Returnerar endast de rader där det finns en matchning i båda tabellerna.*

4. Beskriv kort vad indexering i SQL innebär.

Indexering innebär att skapa en datastruktur (index) för att förbättra sök- och hämtningsprestanda i en databas. Indexering kräver mer underhållning av databasen och behöver uppdateras varje gång nya rader läggs till/tas bort.

5. Beskriv kort vad en vy i SQL är.

En vy i SQL är en virtuell tabell som baseras på resultatet av en SQL-Query. Till skillnad mot "Tables" lagrar vyer (views) inte faktiska data utan representerar endast en vy av data från en eller flera tabeller.

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

En lagrad procedur är en fördefinierad uppsättning SQL-instruktioner som lagras i databasen och kan kallas upp för att utföras och används för att organisera och återanvända SQL-kod.

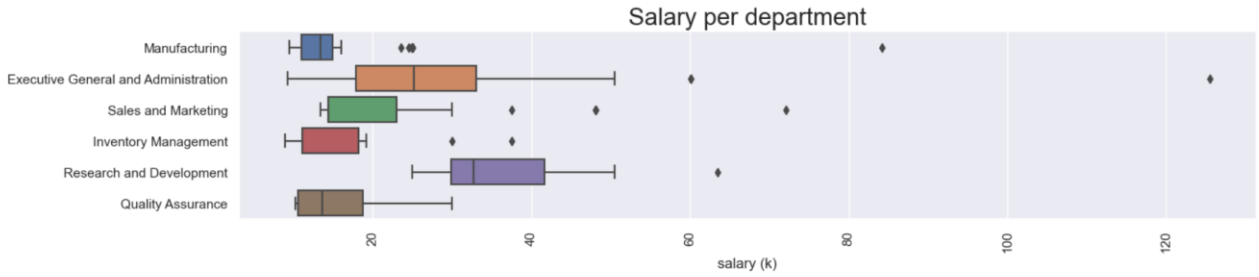
Del 2 - Programmeringsuppgift och Rapport

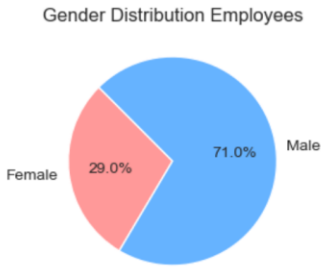
2.1 - Deskriptiv sammanfattning över databasen AdventureWorks2022

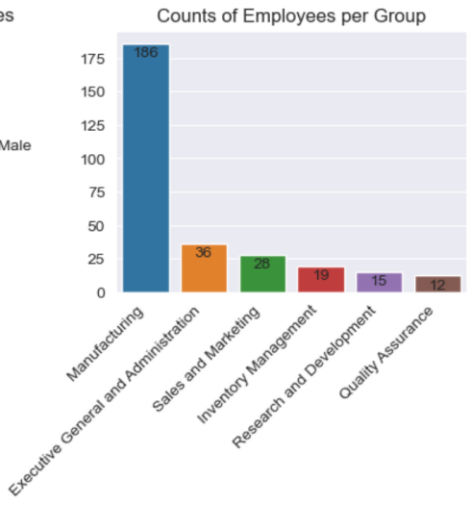
Adventureworks2022 är en databas som sträcker sig från 2011-05-30 till 2014-06-30 över ett fiktivt företag och innehåller cirka 70st olika tabeller inom de 5 olika schemas, "Human Resource", "Person", "Production", "Purchasing" och "Sales".

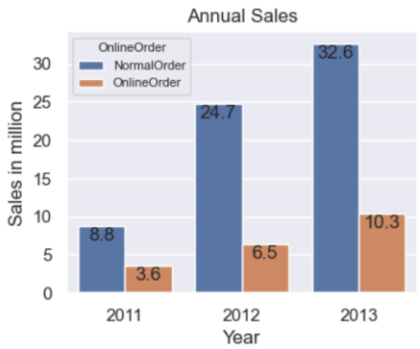
I databasen går det att utläsa att 290 personer jobbar på företaget i en organisationsstruktur innehållandes 5 nivåer. Lönenivån är jämnast blad de anställda som jobbar inom manufactureing, här jobbar också majoriteten av de anställda. Könsfördelningen mellan kvinnor/män är 29% resp. 71 % över hela företaget.

Företagets köper in 295st unika material som antigen säljs vidare eller används som komponenter för att bygga ihop till kompletta cyklar. Företaget har en god årlig försäljningstillväxt och har sedan andra Q2 - 2011 ökat sina försäljningsintäkter inom kategorierna Bikes, Accessories, Clothing och Components från 12 413 872 till 31 201 566 år 2012 och 42 918 447 år 2013 inom. Företaget har sedan haft över 700st unika återförsäljare där cirka 74% av all försäljning sker, Business to Business (BTB). Resterande 26% sker genom Business to Customer genom onlineordrar. Medelordern för BTB och BTC är 21.147USD respektive 1061 USD. Under Q1-Q2 2014 företaget sålt för över 20 miljoner USD.









2.2 - Statistisk Analys

I denna del av rapporten tittar vi närmare på försäljningssiffror, vilka mönster ser vi och vad säger datan från Adventureworks 2022 oss och hur kan vi med hjälp av datan prognosticera Q3 2014? Låt oss observera säljhistoriken från 3 olika perspektiv.

1) fig.1 - Monthly Sales History ser vi att försäljningar varierar från månad till månad. Vi ser en svagt ökande försäljningstrend och att variansen överlag är stor från månad till månad och inga tydliga säsongsvariationer. Särskilt avvikande är försäljningsdatan från 2014 som varierar signifikant. Baserat på datan har försäljningen i princip avstannat helt i juni 2014.

2) I fig.2 - ser vi Monthly Sales Average uppdelat per kvartal. Här får de tidigare kortsiktiga variationer mindre signifikans eftersom vi tittar på rullande medeltal kvartalsvis. Vi ser nu tydligare en försäljningsökning och återigen ingen tydlig säsongsvariation. Här ser vi igen att försäljningssiffror i juni drar ner snittet för det andra kvartalet 2014.

3) fig.3 visualiserar average monthly sales på rullande 12 månadsintervall får att kort och enkelt se hur försäljningsutveckling sett ut sedan 2011. Här ser vi en ganska linjär utveckling men att intäkterna ändå ökat lite mer senaste 12 månader än tidigare period. Vi noterar att försäljningen ökat med cirka 49% från perioden månad 12–24 jämfört mot tidigare period mån 1–12. Under företagets tredje verksamma helår är motsvarande siffra 48%.

Låt oss nu göra ett försök och prognosticera Q3 2014 med tre olika modeller. I denna modellering kommer vi titta på historisk prognostisering. Vi börjar med att anta att datan från juni 2014 är en outlier och potentiellt felaktiga data och exkluderar datapunkten från vår data.

Modell 1: Vi antar att försäljningstillväxten mognat och avstannat helt. Med försäljningshistorik från totalt 12 kvartal (n), medelvärdet ($\bar{x}$) 9,1 miljoner och standardavvikelsen (σ) 2,6 beräknar vi att med 95% konfidensgrad att ***average sales / Quarter*** blir mellan: 7,63 - 10,57miljoner.

Modell 2: Vi räknar med att en årlig försäljningstillväxten på 49% fortsätter vilket motsvarar 12,25% per kvartal men räknar med samma varians, standardavvikelse och 95% konfidensgrad och får ett intervall mellan 9,93 och 12,82 miljoner.

Modell 3: Vi räknar med en årlig försäljningstillväxten på 49% vilket motsvarar 12,25% per kvartal och räknar med en motsvarande ökning av varians och standardavvikelse. Med 95% konfidensgrad får vi ett intervall mellan 9,79 – 13,0 miljoner.

2.3 - Slutsats av statistisk analys

Vid tolkning av datan ser vi tydligt att försäljningsintäkterna (FI) varierar kraftigt från månad till månad. Efter en månad med högre FI följer en månad med markant mindre FI och vice versa. Detta kan bero på beställningsfrekvensen från retailbutikerna. Notera att 74% av all försäljning sker genom retail och att orderstorlekarna skiljer sig enormt mellan BTB och BTC (se bilaga 1 & 2). Mest troligtvis beställer inte retailbutikerna varje månad utan snarare vartannat eller ännu mer sällan. Detta skulle kunna förklara de extrema varianserna, speciellt under 2014 (se bild nedan).

En vidare analys på detta rekommenderas. Förslagsvis kan man titta på de 10 största kundernas beställningshistorik och se om det går att hitta något mönster.

2.4 - Executive Summary

Vi ser stora varianser i sales history vilket om inte förutses kan medföra i att företaget måste ha höga lager och därmed mer risk och högre bundet kapital för att inte bomma en säljorder On-Time in Full (OTIF).

För att inte ha högt lager året runt rekommenderas en mer detaljerad prognos och process för att hantera detta samt att se över safety stocks baserat på prognos. Vidare underlättar detta för planering och minskar kostnader av semester och sjukpersonal i ex. produktion där cirka 64% av personalen arbetar.

Vi rekommenderar även att titta närmare och analysera de 5–10 största kunderna alternativt de kunder som står för 70–80% av FI för att förstå deras beställningsmönster bättre. I vissa fall bör även retailbutikerna kontaktas för att få en bättre förståelse, kanske kan man få information om kommande kampanjer eller till och prognos från butikerna. Upprätta därefter kontinuerlig prognos och jobba med processen för att för en så bra forecast accuracy som möjligt.

Del 3 - Självreflektion

3.1 - Utmaningar under arbetet samt hur du hanterat dem.

Den stora utmaningen var varit relaterat till EDA:n och den statistiska analysen, hur mycket ska man djupdyka i datan för att förstå verksamheten kontra hur mycket ska man djupdyka för att man är nyfiken. Databasen är enorm! Ibland blir man nyfiken och dyker lite för detaljerat för att sedan inse att det inte är relaterat till mitt valda case.

I början var jag mer ostrukturerad tills jag insåg att det inte var hållbart och fokuserade mer på att vara strukturerad. Att Föra anteckningar på relevant data och frågor jag vill ta reda på samt vilka tabeller och vyer som är värda att titta på igen.

Andra utmaningar har varit relaterat till kod hur man ska lösa vissa problem och dubbelkolla så att datan man får fram verkligen stämmer. Det mesta hittar man på nätet hur man kan gå vidare.

3.2 - Vilket betyg anser du att du skall ha och varför.

Jag anser att jag förtjänar VG. Eftersom jag har kunnat visa att jag kan tillämpa samtliga kodfunktioner som har lärts ut i SQL-kursen samt Pythonkod i föregående kurs (DataFrames, Matplotlib och Pandas) men också mer avancerade så som window functions i SQL, Seaborn i Python m.m. Gällande rapportdelen anser jag att den uppnår VG-nivå i hur den är utförd och att mina analyser + reflektioner uppnår VG.

3.3-Tips du hade “gett dig själv” i början av kursen nu när du slutfört den.

Försök vara mer strukturerad ända från början av AdventureWorks2022 projektet. Ha en tydlig plan, hitta ditt business case och håll dig till den, kommentera kod! Var nyfiken på kod och gå in med mind set:et det allra mesta går att lösa på många sätt, experimentera och ha roligt! Brainstorma gärna med andra kurskamrater!

Del 4 - Övrigt & Bilagor

Bilaga 1.1 - Average sales order Offline Order

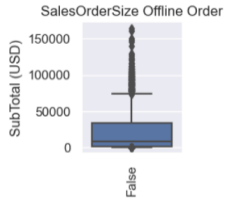
In [53]:
```
sql_SalesOrderHeader = """
SELECT OnlineOrderFlag
    , SubTotal
  FROM [AdventureWorks2022].[Sales].[SalesOrderHeader]
  WHERE OnlineOrderFlag = 0
"""

df10 = pd.read_sql(sql = sql_SalesOrderHeader, con = connection)

# Create Box Plot displaying SalesOrderSize OfflineOrder
sns.set_style = "darkgrid"

g = sns.catplot(x = 'OnlineOrderFlag', y = 'SubTotal', data = df10, kind = 'box', whis=[5,95], height = 2.4, aspect = 1)

plt.title('SalesOrderSize Offline Order')
plt.xlabel("")
plt.ylabel("SubTotal (USD)")
plt.xticks(rotation = 90)
plt.show()
```



In [54]:
```
df10.describe()
```

Out[54]:

	SubTotal
count	3806.000000
mean	21147.583863
std	25559.645345
min	1.374000
25%	1544.970000
50%	8257.084900
75%	34122.384675
max	163930.394300

Bilaga 1.2 - Average sales order Online

In [55]:
```
sql_SalesOrderHeader2 = """
SELECT OnlineOrderFlag
    , SubTotal
  FROM [AdventureWorks2022].[Sales].[SalesOrderHeader]
  WHERE OnlineOrderFlag = 1
"""

df11 = pd.read_sql(sql = sql_SalesOrderHeader2, con = connection)

# Create Box Plot displaying SalesOrderSize OnlineOrder
sns.set_style = "darkgrid"

g = sns.catplot(x = 'OnlineOrderFlag', y = 'SubTotal', data = df11, kind = 'box', whis = [5, 95], palette="GnBu", height = 2.4, a

plt.title('SalesOrderSize Online Order')
plt.xlabel("")
plt.ylabel("SubTotal (USD)")
plt.xticks(rotation = 90)
plt.show()
```

In [56]:
```
df11.describe()
```

Out[56]:

	SubTotal
count	27659.000000
mean	1061.451145
std	1149.026001
min	2.290000
25%	46.470000
50%	594.970000
75%	2181.562500
max	3578.270000